The investigation of code-switching in a computerised corpus of child bilingual language

Catherine Anne Lonngren Sampaio

Submitted to the University of Hertfordshire
in partial fulfilment of the requirements
of the degree of Doctor of Philosophy

September 2014

Acknowledgements

Over the past seven years I have received inspiration, support and encouragement from a great number of individals.

The original motivation to embark on my research journey was sparked by the discovery of the phenomenal work carried out by Brian MacWhinney and colleagues who were responsible for constructing freely available tools designed for transcribing and analysing naturally-occurring spoken data. My corpus, and my study itself, would not exist if it were not for the selfless, inspirational and visionary contribution of this team of researchers.

In terms of support and guidance I would like to thank my supervisor Dr Tim Parke who has been forever calm and encouraging over the many years it took to get to the point of submission of this dissertation. I am also grateful for the comments and pointers provided by my second supervisor, Christina Schelleter. Thanks must also go to the administrative team at the Research Office who have helped ensure that my research journey was as smooth as possible, as I progressed through the different stages of my project.

I am indebted to the staff and pupils at my place of work, Ashtree Primary School and Nursery, who have performed both a direct and indirect role in helping me achieve my research goals. Always supportive, the Head, Beth Kirwan frequently granted me time away from my teaching committments in order to attend research activities and present at conferences. And as for the rest of the lovely staff and pupils at Ashtree, they have helped keep me grounded, providing me with light relief from my studies.

Of course my family deserve special thanks, not least because without my two eldest children, my informants Meggie and James, this study would, literally, not have been possible! And I owe much to my husband who has been so patient over the last seven years and made it possible for me to dedicate so much of my time to my studies (instead of my household duties!) Obrigada por todo o teu apoio e teu amor.

Contents

Abstract

This dissertation describes the investigation of codeswitching in a computerised corpus of child bilingual language, the LOBILL Corpus, which consists of twenty-five hours of recordings of naturalistic interactions between two bilingual Brazilian/English siblings (JAM, 3;6 and MEG, 5;10) and their family members. Collected over three years, the data was transcribed and coded using the CHAT (Codes for the Human Analysis of Transcripts) transcription system developed by MacWhinney and colleagues (MacWhinney, 1991). In addition to standard CHAT coding, language codes were inserted throughout the corpus and a specially developed postcode was added to all bilingual utterances. Addressee information for each utterance was also included.

The longitudinal and heterogenous nature of the corpus and its specific coding allowed for the comprehensive investigation of the children's code-switching practices from both grammatical and pragmatic perspectives. Three levels of analyses were performed using the CLAN (Computerized Language ANalysis) software (ibid). First, quantitative analyses were carried out using the commands FREQ (which outputs frequency word lists), VOCD (which outputs vocabulary diversity scores) and WDLEN (which outputs mean word and utterance lengths). An analysis of the results pointed to the existence of relationships between the various values found and the participatory roles of English and Portuguese in code-switched utterances.

The second level of analysis involved the examination and interpretation of word lists and code lists produced by the use of FREQ. Using Myers-Scotton's 4-Morpheme Model (4-M Model) (Jake & Myers-Scotton, 2009) to interpret the word lists, comparisons of morpheme types revealed the existence of an asymmetry in terms of the contributions of both languages to bilingual utterances. These results were seen to lend support to the Matrix Language/Embedded Language asymmetry proposed in the Matrix Frame Language Model (MFL Model) (ibid). The quantitative analysis of four types of codes (used to code instances of retracings and reformulations, errors, tag questions and metalinguistic usage) provided evidence for the existence of potential relationships between these features of spoken discourse and code-switching.

The third level of analysis allowed further investigation of the above relationships as code-switched utterances were then examined in their linguistic context from a more qualitative perspective. Factors such as geographical location, addressee variables, language development and language awareness were among those seen to account for the variation found in the code-switched data of the siblings and their parents.

The multi-level analysis of the LOBILL Corpus presented in this dissertation shows how corpus-based methodology can contribute to the field of code-switching research. The quantitative investigation revealed relationships which support an existing theoretical model of code-switching (the MFL Model and the supportig 4-M Model). However, it also brought to light how three other types of traditional quantitative measures (vocabulary diversity, word length and utterance length) could be exploited in novel ways in order to characterise the nature of the participating languages in code-switched utterances. Such methodological innovations were translated into a schema which is designed to be used by researchers wishing to interpret such values arising from the analysis of their own code-switched data.

In terms of qualitative analyses, due to its heterogeneous and longitudinal composition, the LOBILL Corpus represents a very rich data set which allows for the examination of several extra-linguistic variables that are known to affect an individual's code-switching practices (as mentioned above).

This study proposes a research methodology which combines the advantages that both quantitative and qualitative approaches to linguistic analysis have to offer. While quantitative analyses have the potential to reveal patterns undetectable by human analysis, qualitative investigation allows for insightful interpretations of such patterns. In the case of this study, the specific coding inserted in the LOBILL Corpus enabled an innovative exploration of the code-switched data. However, it is proposed that such methodological innovations could be exploited in many other lines of linguistic enquiry, leading to new insights which could subsequently lead to the enhancement of existing theoretical models or even the development of new ones.

The final contribution of this research is in terms of the corpus itself. Freely available and transcribed according to internationally recognised conventions, the LOBILL Corpus provides an easily-exploitable, rich data base for others wishing to investigate different aspects of child bilingual language for the pair English/Portuguese.

List of Tables

List of Figures

1. Introduction

The research carried out in this study draws on two major fields of linguistic enquiry, that of bilingualism and corpus linguistics. In my investigation of code-switching (CS) in an electronic corpus of bilingual child language, I propose to marry these two areas by using corpus linguistics methodology to shed new light on the phenomenon of code-switching, specifically that of the code-switching practices of two bilingual siblings.

In order to preface the research undertaken in this dissertation, several topics need to be addressed and it is in Chapter 2 where I review published literature from the field of code-switching and other related fields which will inform both theoretical and methodological considerations of the current study. However, before presenting such a review it is first necessary to tackle the terminology found in the literature on bilingualism and contact phenomena. As is often the case, many terms can carry different definitions depending on the particular research paradigm used to frame each specific study and it is therefore important to first discuss the pertinent terminology and specify the definitions used in this dissertation. I will begin by examining the different views of what it is to be bilingual and then briefly describe language contact phenomena with the aim of arriving at a working definition of code-switching.

1.1 What is it to be bilingual?

In their review of types of bilinguals, Bullock and Toribio (2009) indicate that using the term 'bilingual' is not as simple as it might seem. They write that an individual who is said to have 'native-like control of two languages' (Bloomfield, 1933:56) is often referred to as a 'balanced bilingual', a 'true bilingual' or 'symmetrical bilingual' (Bullock and Toribio, 2009:7). However, they also point out that most researchers would agree that monolingual control over two languages in all aspects of linguistic knowledge and use within all domains is rare, if possible at all (ibid:7). Other types of bilinguals the authors mention include the following: 'Heritage bilinguals', who are second generation bilinguals; 'Second language acquirers' or 'late bilinguals', who they describe as having a linguistic system in place before exposure to second begins; 'naturalistic' or 'folk' bilinguals, who learn a language without formal instruction (for example immigrants and guest-workers); and 'elite' bilinguals, who

have experienced classsroom-based language learning (ibid:8-9). Matras (2010:66) reflects the generally accepted view that whatever type of bilingual a person may be they '(...) - do not, in fact, organize their communication in the form of two "languages" or "linguistic systems." Rather, bilinguals have an enriched and extended repertoire of linguistic structures at their disposal. As part of their linguistic socialization, they learn which word form, construction, or prosody pattern is appropriate in a specific context of interaction.' This notion of 'linguistic socialization', crucial for the present study, will be discussed at greater length in section 2.3.

When it comes to defining societal or individual bilingualism, differences in definition are fewer. However, the problem lies in knowing how to classify individuals who each have a unique linguistic experience, affected in varying degrees by their sociolinguistic and cultural environment. Of particular importance to the present study is what Lanza (2007) calls 'family bilingualism'. The author describes it as 'individual bilingualism within the family in which two languages are spoken'(ibid:10). She shows through her research on family interactional patterns that parental language choices and strategies within the context of the bilingual family have a decisive impact on individual children's code-mixing practices. This concept of 'family bilingualism', of seeing the family as a 'community of practice'(ibid:334) is very important for the current study, as, only by focussing on the discourse data from such a qualitative sociolinguistic perspective will explanations be found for the code-switching practices of the informants.

For researchers studying young bilinguals, an important distinction is usually made between 'infant bilingualism' and 'childhood/successive/sequential bilingualism'. The former involves the simultaneous acquisition of two languages from birth whereas the latter usually involves the establishment of the second language during the school years (ibid:11). Although there has been a suggestion of an age cut-off point of three to distinguish between the two, Lanza says that many researchers avoid using this boundary. For this study on code-switching, both siblings were exposed to Portuguese and English from birth and therefore the question of an age cut-off point is not an issue.

1.2 Language contact phenomena

There is still a great deal of variablity in terms of how different types of contact phenomena can be characterized. This was very recently pointed out by Poplack,

who, when reflecting on over 30 years of research in the field of code-switching since the publication of her seminal paper in 1980[1], says the following:

> "(…) despite 33 years of intense research activity since *Sometimes* was published, there is still no consensus on the nature or identity of even the major manifestations of language contact (codeswitching [CS] and borrowing [B]), let alone the linguistic conditions governing their use. (Poplack, 2013:11)

Most researchers would agree with Poplack that characterising the different manifestations of contact language phenomena remains a challenging task. However, in order to arrive at a working definition of 'code-switching' for this study I need to be able to differentiate it from other contact phenomena. To this end, in the following subsections I will present definitions of the following terms: 'borrowing' ('lexical' and 'nonce'), 'loan translations' or 'calques', 'transfer' and finally 'code-switching'.

1.2.1 Borrowings

Most definitions of 'borrowing' make reference to the notion of 'assimilation' or 'integration' whereby items from the other language 'assume the linguistic structure of a recipient language into which they are incorporated' (Poplack, 2013: p11). Lexical borrowing normally involves morphological and phonological integration of a single lexeme which is fully established in the monolingual lexicon (Bullock & Toribio, 2009:5). For example the word 'self-service' has been borrowed by Brazilian Portuguese to denominate the proliferation of 'self-service' restaurants which cater for the working population at lunch-times. Although this borrowing must have originally occurred via a bilingual speaker, its assimilation has been widespread and the term is now used by monolingual speakers of Portuguese, the majority of whom unaware of its origin or original usage in English. This example illustrates two possible ways of distinguishing code-switches from borrowings: one can look at the degree of assimilation of a form in the community; and one can look at the users of this linguistic form, i.e. where they lie on the continuum of bilingualism (monolingual to 'balanced bilinguals').

For Hickey (2010:18), borrowing does not involve a switch into another language, rather 'Items/structures are copied from language X to language Y, but without speakers of Y shifting to X.' Of course a certain degree of bilingual language

[1] "Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL": Toward a typology of code-switching" (Poplack, 1980).

awareness must have been present in the originator of the borrowing but the final outcome is such that speakers of language Y are not involved in active code-switching or 'shifting'. Therefore, whereas borrowings can potentially be found in the speech of any speaker, code-switches characterise uniquely bilingual behaviour (Bullock & Toribio, 2009:166).

In his discussion of the differences between borrowing and classic code-switching, Winford (2010:182) concludes that lexical switches and lexical borrowings should be treated as manifestations of the same underlying process of borrowing, that 'differences among the outcomes have to do with the degree to which the processes apply, and the extent to which the switches become conventionalized as fixed lexical selections' (ibid:183). In other words, the linguistic outcomes are essentially determined by 'social convention'(ibid:183).

When we examine the motivations behind borrowing, the parallels between the phenomena of borrowing and code-switching clearly converge. Social motivations for the borrowing of overt elements may include the naming of a concept for which there is no term in the Recipent Language (RL)(Winford, 2010:177) but may also be due to the prestige of the dominant language, leading to the preference for the borrowed term despite there being an equivalent in the RL. Sociolinguistic and sociopolitical factors affecting the amount of borrowing occurring in a community are manifold and may include the patterns of social interaction between the groups in question, the degree of bilingualism, demograhics and power relationships, attitudes towards the languages, language loyalty and language ideology (ibid:178). Linguistic factors affecting, or constraining, borrowing include the degree of typological distance between the languages in contact, with greater congruence facilitating borrowing (ibid:178).

All of the above-mentioned motivations can equally be said to affect the code-switching practices of an individual or community. It appears, then, that the only way to effectively distinguish a case of borrowing from a code-switch is to be able to determine its degree of assimilation into the recipient language. This means that the researcher has to necessarily look beyond a purely linguistic analysis of foreign material in the speech of a bilingual. Only by examining the sociolinguistic practices of the community to which a particular bilingual speaker belongs, will it be possible to establish if that individual is employing a borrowing or a code-switch.

Lying on the 'assimilation' continuum one also finds 'nonce' borrowings which 'can occur spontaneously in the speech of bilinguals, blurring any boundary that can be drawn between these contact forms on structural criteria alone'.(Winford, 2010:178). These types of on-the-spot borrowings are ones which are morphologically, syntactically and phonologically integrated into the recipient-language lexicon (Poplack, 2001) but have not been widely propagated and therefore have not become assimilated into the language of the community (Matras, 2009:106).

1.2.2 Loan translations or calques

Whereas borrowings involve the copying of foreign morphemes into a recipient language, 'loan translations' or 'calques' involve the importation of foreign patterns or meanings while retaining native-language morphemes (Bullock & Toribio, 2009:5). For example, the word for 'skyscraper' in Brazilian Portuguese is 'arranha-ceu' which literally translates as 'scratch-sky'. As in this example most cases involve partial translation (Backus & Dorleijn, 2009:76). In order to differentiate loan translations from the phenomena of 'interference/transference' (see 1.2.3), Backus and Dorleijn point out that while the former is usually restricted to the translation of specific expressions, the latter involves the copying of general grammatical structure (ibid:78).

1.2.3 Transfer

Put simply, 'transfer' can be described as the influence of one language on another. Also known as 'interference', 'cross-linguistic influence', 'convergence', 'intersystemic influence', 'substrate'/'superstrate'/ 'adstrate influence' (Treffers-Daller, 2009b:58), Jarvis elaborates further in his definition of 'transfer' and describes it as:

> "the influence that a person's knowledge of one language has on that person's recognition, interpretation, processing, storage and production of words in another language"(Jarvis, 2009: 99).

This type of transfer appears to be manifest in the following example from my corpus data where JAM is talking to his mother:

(1)

*JAM: <Jake@pn <he's got>[*] five years old>[@en] .
%add: MOT F055: L132[2]

Although only using English morphemes, JAM uses '(ha)s[3] got' instead of the expected 'is'. It does seem that his knowledge of Portuguese (where age is expressed through the verb 'ter', the equivalent of 'to have') has influenced the production of this particular utterance in English. The fact that his next utterance shows the correct use of 'is' indicates that this might be a case of what Paradis terms 'dynamic interference', a performance error where an element of one language appears in the sequence of another language *inadvertently* (my emphasis) (in Bullock & Toribio, 2009:61). This contrasts with 'static interference' where an element becomes part of the implicit grammar of an individual (ibid). If, over time, Portuguese were to become increasingly dominant, it might be that this type of performance error would appear more and more frequently in JAM's English until at some point it became embedded in his grammar.

Although the utterance discussed above may indeed be a manifestation of transfer, Jarvis and Pavlenko (2008) argue that such confirmation would only be possible after an examination of the distribution of this particular structure in the source language and in the language of other bilinguals or L2 learners with different L1s. For further details on how to indentify transfer in bilingual data, readers can consult the above-mentioned volume, especially Chapter 2 (ibid, pp.27-60) where the authors discuss the issues and challenges related to the identification of transfer.

1.2.4 Code-switching

Through the discussion of the terms above, which all describe phenomena resulting from language contact, it will now be possible to draw out those aspects which characterise code-switching and differentiate it from 'borrowings', 'calques', and 'transfer'. The aim here is to arrive at a working definition of code-switching before launching into a more detailed review of the body of research on code-switching (in Chapter 2).

Let us consider the following definition given by Jake and Myers-Scotton:

[2]The format used for indicating the source of examples from the LOBILL Corpus is Fxxx: Lxxx, where F is followed by the file number and L is followed by the line number. While the File List can be found in Appendix A, details on the transcription and coding system of the corpus can be found in section 3.2 and in Appendix B.

[3] Although one could argue that 'he's' could also be the abbreviation of 'he is', it is the use of 'got' that is key here.

> ‘Codeswitching (CS) refers to language use that consists of material from two or more language varieties at any level from the discourse to the clause.’
> (Jake & Myers-Scotton, 2009:207)

If we recall that ‘borrowings’, ‘calques’ and ‘transfer’ all involve language use that consists of ‘material’ from two or more languages, it appears that this definition needs expanding in order to apply exclusively to code-switching. Firstly, while ‘calques’ and ‘transfer’ do indeed involve borrowing material from one language and inserting it into another language, this material consists of the importation of meaning and/or underlying structure which is ‘translated’ into the borrowing language. That is, all the surface-level morphemes come from the same language. Code-switching, in contrast, involves the juxtaposition of surface-level morphemes from two different languages or language varieties. The word ‘material’ thus needs further specifciation (see below). In order to exclude ‘borrowings’ from Jake and Myers-Scotton’s definition, we would have to specify what we mean by ‘language use’. As was seen above in 1.2.1, although ‘borrowings’ involve the insertion of non-native morphemes into a recipient language, this material has been assimilated into the language practices of the wider community in such a way as to become part of the monolingual lexicon. Over time it has become part of the established, or ‘static’, repertoire of a community’s language use. This is in contrast to a code-switch which usually reflects more dynamic language use: it is spontaneous, individualistic and novel.

I would like to propose that Jake and Myers-Scotton’s definition would therefore benefit from the following two additions:

> ‘Codeswitching (CS) refers to dynamic language use that consists of surface-level material from two or more language varieties at any level from the discourse to the clause.’

It is important to point out that implicit in this definition is the idea that the 'surface-level' material retains its original 'donar language' identity and is not integrated into the 'recipient language' (Poplack, 2013:11). As seen above in 1.2.1 this contrasts with borrowing where, in most cases, the borrowed items are phonologically, morphosyntactically and syntactically integrated into the receiving language.

In this study, then, a working definition of code-switching is that it is a dynamic linguistic process whereby a bilingual speaker juxtaposes (intentionally or unintentionally) surface-level material from two or more languages within a phrase or across utterances.

Before leaving this discussion on the definition of code-switching, it is important to mention two other terms which are often used interchangeably with code-switching in the literature. These are 'language mixing' or 'code-mixing'.

For Lanza, who studied infant bilingualism, 'language mixing' is a generic term for any type of linguistic interaction between two or more languages and she views CS as a type of language mixing (Lanza, 2007:3). In Yip and Mathews' study on infant bilingualism (2007) the preferred term is 'code-mixing'. Indeed in such studies on 'early mixing' researchers appear to make use of different terminology to reflect their view that 'language mixing' or 'code-mixing' in young children is essentially different to adult 'code-switching'. Cantone contests this assumption and prefers to use the term 'code-switching' for both cases in order to support her view that as the same constraints can be seen to operate in both infant and adult bilingual speech, there is no real difference between CS as it occurs in adult speech and in infant speech (Cantone, 2007). In section 2.2.3 the studies which provide evidence for and against the need to differentiate infant and child 'code-switching' from that of adults will be dealt with in greater detail.

Even within research on adult bilingual speech there are some researchers who prefer to use the term 'code-mixing' instead of code-switching. Perhaps the most influential of these is Pieter Muysken who uses the term 'code-mixing' to refer generally to "all cases where lexical items and grammatical features from two languages appear in one sentence" (Muysken, 2000:1) and considers the term 'code-switching' only appropriate for what he describes as the 'alternational type of mixing' (see section 2.2) as it "suggests something like alternation (as opposed to insertion)"(ibid:4).

Although Muysken's typology of 'code-mixing' will be described and referred to in my research, it is important to highlight here that I will be using the term 'code-switching' as an all encompassing term instead of the term 'code-mixing'. One of the reasons for this is its currency of use in the literature. More importantly, however, is the fact that the term 'code-mixing' can conjure up rather negative images of speakers who are unable to keep their languages separate, resulting in involuntary 'mixing' which is often seen as undesirable. Although speakers may not always be aware of when they are using two languages in an utterance, this might simply reflect the sociolinguistic norm of a certain community or family practice. The more neutral term, 'code-switching', will therefore be used throughout the thesis, 'code-mixing'

only being maintained (in italics) when directly referring to the work of particular authors, such as Muysken (see next chapter).

1.3 Rationale in brief

The variability in the use of terminology in the field of code-switching appears to be a reflection of the lack of consensus as to the 'nature or identity' of the phenomenon itself (see Poplack's earlier quote at the beginning of this section). Clearly there is a need for more studies on code-switching but Poplack states that in order for progress to be made in this field, future research should focus on *quantitative* analyses of code-switches in corpora of spontaneous speech. Only by taking into account what Poplack terms the 'principle of accountability' (Poplack, 2013:11) will it be possible to try and make sense of the inherent variability found in bilingual discourse. The importance of being able to analyse naturally occurring bilingual data is being increasingly recognized as it affords insights that elicted data cannot offer (Travis & Cacoullos, 2013). This evidently calls for the compliation of corpora of spontaneous bilingual speech and in section 2.2.4 we will see the progress that has been made in this area.

The current study seeks to address the clear need for more quantitative studies of code-switching in corpora of naturally-occurring bilingual discourse. However, as will be seen throughout the dissertation, the contribution of my investigation of code-switching in a corpus of child bilingual language goes beyond that of providing original results. There is also originality in terms of methodology, and the corpus itself represents a significant contribution to a research field in great need of readily available bilingual data.

The aim of this introduction has been to clarify the object of my investigation and outline a brief rationale for this study. In the following chapter I will now provide a review of the research carried out in the field of code-switching (2.1) and discuss the different methodologies used to investigate this phenomenon (2.2). This will then allow me to specifiy the research questions that will guide my investigation (2.3).

2. Literature Review

Having addressed more general terminological issues relating to bilingualism and language contact phenomena, I am now in a position to examine the literature on code-switching. Firstly, both grammatical and sociolinguistic treatments of code-switching will be discussed and CS research in bilingual children will be highlighted, along with the key notion of language socialization and its relevance to this study. Then, after presenting existing proposals for a more holistic approach to CS research, developments in the study of code-switching in methodological terms will be outlined. This will necessarily include a look at the computerised bilingual corpora already available for research purposes. I will end the chapter with a discussion of the rationale behind my study and highlight the particular research questions I am seeking to answer.

2.1 Research in code-switching

In this section, both grammatical and sociolinguistic treatments of code-switching will be discussed in order to tease out the aspects of code-switching that will be the focus of analysis in this study. Firstly, influential studies focussing on the grammatical properties of code-switching will be reviewed and then research into the social dimensions of code-switching will be examined. It is relevant here to mention that grammatical treatments mostly look at 'insertional' or 'intra-sentential' code-switching, which involves the use of surface-level morphemes from two languages *within* a phrase or utterance. This contrasts with 'alternational' or 'inter-sentential' codeswitching which involves switching *between* phrases or utterances. Both types can be found in sociolinguistic studies of code-switching.

2.1.1 Grammatical approaches to code-switching

Ever since the 1970s, many researchers have focussed their energy on attempting to formulate what have been termed as 'constraints' on code-switching. That is, what can and cannot happen when two languages interact within a phrase or utterance (i.e insertional code-switching). Over the years, many of these proposals for universally applicable constraints have been refuted following the emergence of more and more contradictory empirical evidence. For a summary of these (mostly foiled) attempts to explain code-switching behaviour, the reader can refer to MacSwan (1997) and his

discussion of the proposals put forward by Poplack (1980, 1981), Joshi (1985), Di Sciullo, Muysken and Singh (1986), Mahootian (1993) and Belazi, Rubin and Toribio (1994). In this discussion MacSwan shows, through the presentation of empirical data and his own counter-examples, that all of the constraints are subject to criticisms, some being more flawed than others. However, he does point out that they have led to insights which have allowed for progress in our current understanding of code-switching behaviour.

Rather than review the proposals mentioned above, the discussion in this section will focus on current proposals and models which claim to account for and predict grammatical code-switching. Presently the two opposing camps which appear to represent the current division over how best to account for the variety found in CS data are formed by Myers-Scotton and her team (Jake and Gross) and MacSwan and his supporters (in particular Cantone). Through the ever increasingly complex Matrix Language Frame (MLF) model, Myers-Scotton and Jake aim to predict all possible forms of CS speech (Myers-Scotton, 2002; Myers-Scotton & Jake, 2009). MacSwan, on the other hand, aims is to show that, from a Minimalist approach, a model developed exclusively for CS (such as the MLF model) is completely superfluous as "Nothing constrains codeswitching apart from the requirements of the mixed grammars"(MacSwan, 2005a:5).

Before providing more details about these two opposing models which claim to account for the variation in CS, it is important to note that it is not the aim of this study to provide a detailed grammatical analysis of the data in the corpus, such as that provided by Myers-Scotton and MacSwan in their research. As will be seen, sociolinguistic aspects of CS are considered to be more central for the present research. However, certain notions and terminological issues surrounding the grammatical treatment of CS need to be addressed in order to provide a descriptive framework for the analysis of the data at hand. This can only be done through a discussion of key aspects of the models mentioned above.

2.1.1.1 The Matrix Language Frame Model (Myers-Scotton)

First proposed in 1993 (Myers-Scotton, 1993a), the Matrix Language Frame Model (MLF) has been very influential in the field of CS and along with the subsequent 4-M model of morpheme classification, much of the terminology used to discuss current insertional CS data is drawn from these models.

There are two basic principles underlying the MLF Model: the Asymmetry Principle (Jake & Myers-Scotton, 2009:209) and the Uniform Structure Principle (ibid: 210). According to the former principle, there is always asymmetry between the two (or more) languages participating in CS clauses and this is reflected in the data in two main ways. Firstly, the abstract morphosyntactic frame of the bilingual clause largely, or entirely comes from one of the languages (the System Morpheme Principle): Myers-Scotton names this language the Matrix Language (ML) while the other participating language is called the Embedded Language (EL). Secondly, the word order of a bilingual clause tends to follow that of one of the languages (the Morpheme Order Principle)(ibid:208). This asymmetry in the roles of the participating languages is evident in diverse CS corpora and leads to Myers-Scotton's second related principle underlying the MLF Model, the Uniform Structure Principle, which she states as follows:

> 'A given constituent type in any language has a uniform abstract structure and the requirements of well-formedness for this constituent type must be observed whenever the constituent appears. In bilingual speech, the structures of the Matrix Language are always preferred...'(Myers-Scotton, 2002:8).

Although this means that basic clause structure is uniformly provided by the ML, the Embedded Language may provide grammatical structure in the form of 'Embedded Language islands'. These islands are 'full constituents consisting only of Embedded Language morphemes occurring in a bilingual CP that is otherwise framed by the Matrix Language'(ibid:139). Typical EL islands are often of a formulaic nature such as idioms, set collocations, adverbial phrases of time or place.

A key element to understanding the MLF Model and its underlying principles is the distinction Myers-Scotton makes between the different types of morphemes that are supplied by the ML and the EL in bilingual clauses. This is captured in the supporting 4-M model which classifies all morphemes into four different types based on their roles in phrase or clause structure:

> 'The model's classification is based on how morphemes differ from each other in whether they are meaningful and therefore are primarily called by speakers' intentions, or whether they primarily build grammatical structure. Thus, the basic division in the 4-M model is at an abstract level, between conceptually-activated vs. structurally-assigned morphemes' (Jake & Myers-Scotton, 2009: 214)

The two types of conceptually-activated morphemes are called 'content morphemes' and 'early system morphemes' and their role is to convey semantic content (ibid:215). Whereas content morphemes include nouns, verbs, adjectives and some prepositions, early system morphemes include plural affixes, many determiners and verbal prepositions. The latter supply semantic meaning but cannot occur on their own. Structurally-assigned morphemes are termed 'late system morphemes' and can be one of two types, 'bridges' and 'outsiders'. As the name indicates 'bridge late system morphemes' join smaller constituents to construct larger constituents, typical examples being genitive or partitive constructions. The role of 'Outsider system morphemes' is to co-index relationships and make argument structure more transparent (ibid:216). For example, through subject-verb agreement the relation between subject NPs and inflected verbs is co-indexed.

In order to make more sense of this terminology I have reproduced the following figure which Jake and Myers-Scotton uses to summarize the hierarchy of distribution of morphemes based on the 4-M model (ibid:219):

Figure 1. Distribution of morpheme type in bilingual constituents (Jake & Myers-Scotton, 2009)

Content Morphemes From the ML or EL	**Early SMs** More from the ML than the EL	**Bridge SMs** Rarely from the EL	**Outsider SMs** None: only in monolingual EL consituents

The finding that only the ML provides outsider morphemes in CS, has led Myers-Scotton to formulate a hypothesis that proposes to explain this difference in distribution morpheme type. Called the 'Differential Access Hypothesis', it attributes this differential distribution to patterns of accessibility in language production:

> 'The Differential Access Hypothesis: The different types of morpheme under the 4-M model are differentially accessed in the abstract levels of the production process. Specifically, content morphemes and early system morphemes are accessed at the level of the mental lexicon, but late system morphemes do not become salient until the level of the formulator.'(Jake & Myers-Scotton, 2009:218)

It is at the formulator level that the conceptually-activated lexical entries (content morphemes and early system morphemes) are then assembled into larger constituents, requiring the activation of late system morphemes (bridges and outsiders). The directions which come from the formulator are language-specific, that

is they all come from one language, the Matrix Language. The resulting surface level output, therefore, contains the morphosyntactic patterns of the ML and those of the EL in the case of EL islands.

It is important to note that Myers-Scotton's classification of morpheme types in the 4-M model is not a classification of lexical categories. That is, a morpheme type does not correspond necessarily to a lexical category. Jake and Myers-Scotton illustrate this in their research on prepositions in CS, showing that while some prepositions, such as in '*from* someone' or '*beside* the road' function as content morphemes, prepositions in phrasal verbs such as 'throw *away*' are early system morphemes. The authors also exemplify those prepositions which can be classified as bridge and outsider system morphemes (see Jake & Myers-Scotton, 2009:218-226 for discussion).

Apart from the variation in morpheme types which may exist within a particular lexical category *within* a language, cross-linguistically these lexical categories may also have different grammatical properties. This was also pointed out by Gardner-Chloros who discusses how the distinction between function words and content words is not always clear-cut and does not occur in the same way across all languages (Gardner-Chloros, 2009:102-4). She uses this fact as an arguement against the wide-spread use of the term Matrix Language, stating that such indefinition makes it problematic to clearly define what a Matrix Language is (ibid:203).

Indeed, trying to establish the ML of a bilingual clause based on lexical categories, rather than morpheme types, could be counter-productive. However, Myers-Scotton's MLF model and the supporting 4-M model require an analysis of CS which uses morpheme types, and not lexical categories, to establish the roles of the participating languages. And if, as she and Jake state, 'the definitions of the morpheme types are universal'(Jake & Myers-Scotton, 2009:219), this means that the author's models should be applicable to all classic CS data, both explaining and predicting what does and does not occur in any language pair.

Myers-Scotton and Jake (ibid:239) recognize the existence of other approaches to CS, mentioning specifically that proposed by MacSwan (see next section). Indeed it is side by side in the very same volume (*Multidisciplinary Approaches to Code Switching,* edited by Isurin et al, 2009), that we find arguments for these opposing approaches put forward first by Myers-Scotton (& Jake) in

Chapter 9 and then by MacSwan (& Cantone) in Chapter 10. In both chapters we find criticisms of the opposing approach and arguments which serve to try and convince the reader to align with one of the sides. It appears to be an 'either-or' situation, apparently with no compromise or middle ground.

To allow for neutrality as a starting point for this study, the next section will now outline the approach to CS espoused by MacSwan.

2.1.1.2 The Minimalist Approach (MacSwan)

As mentioned in the introductory part to this section (2.1.1) many grammatical treatments of CS have as their ultimate goal the search for universal constraints which are specifically applicable to CS data (bilingual data). According to MacSwan, these constraints are unnecessary as '...all the facts of code-switching may be explained just in terms of principles and requirements of the specific grammars used in each case, including principles and requirements of Universal Grammar'(MacSwan, 2005b:69). His approach has a theoretical framework based on the latest developments in generative grammar (Chomsky, 1991, 1994) which now propose a lexicalist theory of grammatical structure. Called the 'Minimalist Program', it is a theory which states that grammatical structure is projected by lexical items. That is, parameters, or rules, are encoded in the lexicon rather than there being an independent system of syntactic rules which govern grammatical structure (see earlier generative theories in Chomsky 1957, 1965, 1970 and 1981). MacSwan has taken this Minimalist Program and simply applied it to bilingual language, stating the following:

> 'If all syntactic variation is associated with the lexicon, as in the Minimalist Program, then CS may be seen as the simple consequence of mixing items from multiple lexicons in the course of a derivation.'(Cantone & MacSwan, 2009:251)

In his model of CS (MacSwan, 1999, 2000), lexical items can be drawn from the lexicon of two or more languages, bringing with them encoded features which must then be checked for convergence. This checking of features occurs in the same way for both bilingual production and monolingual production, 'with no CS-specific mechanisms permitted' (Cantone & MacSwan, 2009:251). As grammatical requirements are encoded in the lexical items of each language, the resulting theory of CS can be stated very simply:

> Nothing constrains CS apart from the requirements of the mixed grammars (MacSwan, 1999).

Clearly this theory is in stark contrast to the constraint-oriented MLF model proposed by Myers-Scotton which was discussed in detail in section 2.1.1.1. MacSwan and Cantone devote over four pages of their paper to discussing the weaknesses of the MLF model as a theory of CS, analysing data which appear to contradict the principles underlying the model. They criticise a further amendment to the MLF model which allows the existence of 'internal EL islands' as well as just EL islands. These 'internal EL islands' are described by Myers-Scotton and colleagues as being a constituent consisting of EL morphemes in the EL order but smaller than a maximal projection (Jake, Myers-Scotton & Gross, 2002:76). The result of this and other amendments, according to MacSwan & Cantone, is the sanctioning of any and all CS examples (Cantone & MacSwan, 2009:254).

Indeed it does appear that ever since the MLF model was first proposed by Myers-Scotton in 1993, it has undergone further and further amendments in order to deal with the contrary data as it emerged. The development of further principles and accompanying hypotheses have certainly served to expand the MLF and 4-M models However, has this desire to be able to account for all CS data actually affected their explanatory and predictive power? If, as MacSwan and Cantone point out, the models appear to have reached the point of essentially permitting all CS data, how useful, in fact, are they for researchers wishing to analyse their own data?

MacSwan's Minimalist theory of CS has the historical backing of other researchers who also believed that a theory of CS without specific CS constraints, or without a third grammar, was preferable. These include Pfaff (1979), Poplack (1981), Woolford (1983), Lipski (1985), di Sciullo, Muysken and Singh (1986) and Belazi, Rubin and Toribio (1994). Earlier generative approaches had posited that lexical insertion only occurred after word order had been laid out (Cantone & MacSwan, 2009:256) and this had posed a problem because evidence showed that the language of the lexical item seemed to play a role in the structure in which it occurred. In current generative theory, this no longer presents a problem. With the advent of the Minimalist Program, where lexical insertion dictates structure, a constraint-free theory of CS could now be implemented:

> 'Within the Minimalist Program, structures are built from a stock of lexical items, essentially beginning with lexical insertion (formalized as Select). This

> important development permits CS researchers to probe the structural consequences of particular lexical items from specific languages, with no need to keep track of which languages may contribute which specific lexical items during a final stage of lexical insertion.'(Cantone & MacSwan, 2009: 256)

Thus, for MacSwan and colleagues, it is the study of the 'structural consequences' of lexical items from different languages which should be the central focus of CS research, and not the search for constraints beyond those specified by the monolingual grammars.

The Mimimalist Approach reflects the desire to achieve parsimony by presenting a theory of CS which can be explained using an already existing monolingual model. However, this appears to imply the position that bilinguals are simply two monolinguals in one. Research in bilingualism has shown that this view is much too simplistic, with evidence pointing to the underlying importance of considering such issues as language dominance and how they affect bilingual output. Myers-Scotton and Jake argue that 'it makes sense that bilingual data present a complication not found in monolingual data; that is two systems of grammar are in contact.'(Jake & Myers-Scotton, 2009:239). They go on to say that even if it can be assumed that the same universal grammatical principles are in operation for each language, the question of which language contributes 'the language-specific parts of the grammar of a bilingual clause'(ibid:239) still remains. According to Myers-Scotton and Jake, empirical evidence supports a theory of unequal participation of the two languages and this asymmetry in the roles of the participating languages and those of the different morpheme types needs to be included in any theoretical model which aims to account for the variation in CS.

Thus we have two opposing theoretical models which both claim to be able to explain and predict CS: the ever-increasingly complex models developed by Myers-Scotton and colleagues; and the rather simplistic model proposed by MacSwan. It is at this point that one might wish to reflect upon the very fact that there are such discrepancies in grammatical approaches to CS despite over thirty years of research in this area. As more and more empirical evidence becomes available for analysis it appears to me that such models will struggle to account for all the variation which is found in code-switched data.

This calls into question the usefulness of such models in CS research and highlights the need to perhaps focus on non-grammatical factors affecting CS.

Indeed in recent work Gardner-Chloros has called for this change in focus, arguing that there is no single set of grammatical rules that can account for the variation that exists in CS and that it cannot be described as if it were made up of the combination of two systems outside the individual (Gardner-Chloros, 2009:173). In her view, CS can be of a highly personal nature, individuals constructing 'their own systems from the input and models to which they are exposed' (ibid). This necessarily involves focussing on more pragmatic and sociolinguistic dimensions of CS and it is these issues that this review will now consider in the following section.

2.1.2 Sociolinguistic and pragmatic approaches to code-switching

In contrast to grammatical approaches to CS, sociolinguistic and pragmatic studies of CS share the common goal of searching for and analysing the meaning brought about by CS, whether through examining social factors or through more local conversational-level factors. As will be seen in this section, although the focus of these two approaches may indicate that explanations for code choice can simply be divided into those related to social factors and those related to discursive functions, in reality the motivations behind CS are frequently multiple and complex, as Matras comments: ‘The meanings that codes acquire can be multidimensional and can draw on their macro-social functions, on their significance for individuals, and on the choices that have already been made earlier in the same interaction.’(Matras, 2009: 124).

Due to this overlap and interplay of motivations behind CS, the discussion in this section will not be neatly divided into two sub-sections. Rather, the discussion will reflect the interwoven nature of the multiple factors affecting CS, which are, at times, difficult to tease apart.

In speaking about language contact generally, Thomason and Kaufman (1988:35) came to the conclusion that sociolinguistic factors were more important than structural, or grammatical, factors in determining the outcome of languages in contact. Researchers focussing more specifically on code-switching have also found evidence that supports the idea that ‘sociolinguistic factors are the key to understanding why codeswitching takes the form it does in each individual case.’(Gardner-Chloros, 2009:41). When comparing the CS practices of two different groups, a Punjabi/English and a Greek/English group, Gardner-Chloros (1997:270) found that, despite the latter language pair being more typologically closer than the

former pair, more CS was actually found in the Punjabi/English-speaking group. In this case, sociolinguistic factors appear to override the structural closeness of the languages, giving results which contradict the posited notion that the closer the languages are typologically, the more likely CS is to occur (see Poplack's equivalence constraint, Sankoff and Poplack, 1981:4).

From a sociolinguistic point of view, Gardner-Chloros (2009:42) points out that there are three types of factors which may affect the form of any particular instance of CS: firstly, community-level factors which are independent of the individual such as prestige, power relations and the associations of each language variety with a particular context or way of life; secondly, individual-level factors which include competence, social networks and relationships, attitudes and ideologies, self-perception and the perception of others; and thirdly, conversation-level factors which involve the use of CS as a tool to structure discourse.

Sociolinguistic approaches effectively necessitate ethnographic knowledge of individual speakers within their family and community contexts. Detailed studies such as Blom and Gumperz's (1972), Stroud's (1992 and 1998) and Zentella's (1997) reveal how central such ethnographic knowledge is for a multilevel explanation of CS practices, Stroud going so far as to say that only a deeply ethnographic approach can get anywhere near understanding the 'meaning' of CS from an emic perspective (in Gardner-Chloros, 2009:77).

There have been attempts to offer predictive rules for CS based on social conventions, such as the Model of Markedness devised by Myers-Scotton (Myers-Scotton, 1993b). This model proposes that far from being random, language choices are predictable via a set of indicators associated with each language, the default language choice of the community known as the 'unmarked' language. 'Marked' choices would be the 'unexpected' use of a language in a conversation. Matras questions the validity of such a model asking 'But if unmarked code choice is predictable through a set of social conventions, why do speakers defy these very same conventions and make marked choices?'(Matras, 2009:116). It is evident that the creative and dynamic nature of CS means that the code choice of a particular speaker in a given conversation cannot be explained using soley social meanings as parameters. In Stroud's view 'meaning' is a "negotiated product" which emerges from the conversation and this creates a problem when trying to assign meaning to CS, begging the question "Whose meaning is it?" (Stroud, 1992:151).

The notion of a 'trigger' developed by Clyne (1967) is useful in helping us understand that CS is indeed responsive to events surrounding communicative interaction (Matras, 2009:114). An alternational switch, where the speaker code-switches and continues in that language for the rest of the turn, can be triggered by subtle changes such as the appearance of certain words, shifts in topic, inclusion or exclusion of participants or the presence of bystanders (ibid:114). Although a more pragmatic approach to CS needs to take into account the social roles of languages, it recognises that the motivations to choose one language over another are multiple and complex, each language representing 'a whole array of functions and symbolisms at a given moment in conversation.'(ibid:115). Moreover, as Matras shows, the markedness of code selection can in itself be dependent on the context and created and negotiated by the participants (ibid:23-4). For example, in the analysis of a spoken narrative (ibid:121), he found that the speaker used the language of the initiator (the interviewer) to tell the narrative and code-switched to the other language in order to convey evaluations, attitudes and justifications in the form of side-comments. Matras points out that had the initiator started in the other language, the roles of both languages would probably have been reversed. In this case CS served as a tool to structure the narrative discourse.

Much research has been carried out on the use of CS as a tool to structure discourse. Researchers using a Conversation Analysis approach (Auer, 1995: Li Wei, 2005), while recognising an indirect link to the social roles of the participating languages, claim that the crucial driving force behind the choices that a speaker makes is provided by the local goal of the interaction (Auer,1995 in Matras, 2009:121). Such conversation-oriented functions include the following: the highlighting of reported speech, the use of parenthesis or side-comments, reiterations or quasi translations for emphasis, a change of mode (e.g. from formal interview to informal conversation), language play and topicalization (focus or contrast)(Matras, 2009:116-117). It is important to point out that such discourse functions can be perfomed monolingually, often by a change in tone. However, through the use of a code-switch, these strategies become more salient as they are marked twice over (Gardner-Chloros, 2009:77).

Although a pragmatic approach to CS may involve attempts to systematize code-switching practices, there is an awareness of the fact that these attempts at constructing systems may prove frustrating as 'speakers can then turn round and

deliberately ignore them or subvert them, in their online productions, for their own communicative ends.'(Gardner-Chloros, 2009:88).

Code-switching can also serve as a way of accommodating an interlocutor's linguistic preferences or competences and a study carried out by Woolard (1997) showed that there may be a relationship between this type of linguistic accommodation and gender: she found that when talking to Castillian friends, Catalan adolescent girls demonstrated greater tendency towards linguistic accommodation to their interlocutors than the Catalan boys.

Attitudes also have an important role to play in determining an individual's CS behaviour and these can be influenced by the community, school and family practices. Formal and informal language policies will have a big impact on the degree of acceptability of CS within a community or family environment. This is especially so for children growing up bilingually: their use of CS will be greatly influenced by the language practices of those around them as they become 'socialized into language and through language'(Ochs & Schiffelin, 1984, 1995). For the present research, the notion of 'language socialization' is crucial and for this reason will be discussed at length in section 2.1.3.1 after briefly focussing on code-switching research specifically concerning children.

2.1.3 Code-switching research in bilingual children

In the discussion on terminological issues surrounding code-switching (in 1.2.6) it was pointed out that the frequent use of the term 'code-mixing' as opposed to 'code-switching' in studies on infant bilingualism reflects the commonly-held assumption that adult CS is different in nature to child *code-mixing*. That is, that the alternation of languages in infant speech is not yet constrained in the same way as in adult speech. In this section evidence for and against this view will be presented.

The debate on whether constraints exist in infant CS centres on bilingual children whose languages are in the developmental stages. This normally includes the study of infants up to the age of 3, which is when their language becomes more adult-like in terms of structural properties. A hypothesis put forward by Meisel (1994), named 'The Grammatical Deficiency Hypothesis' claimed that there is a stage in language development in which a child's word combinations are not constrained by principles of grammar (not yet evolved and in evidence), and thus that language mixing at this stage is not constrained by structural principles either. This implies that

there is a change at some point from unconstrained CS to constrained CS. This change was not corroborated by Paradis et al (2000) who found evidence of CS structural constraints before the use of INFL-related morphology occurs in both languages. Based on this evidence Meisel's hypothesis is not supported.

Bernardini and Schlyter (2004) studied the simultaneous acquisition of Swedish and Italian and propose that unbalanced bilingual children use the more developed language in order to build sentences in the weaker language; that is, it is a particular stage in language development. However, Muller and Cantone (2009: 206) argue that the fact that bilingual children's mixing patterns are so varied (some mix a lot in both languages, some only in one, some not at all) and not necessarily uni-directional is evidence for the idea that language mixing appears to be more of an individual choice than a developmental stage.

In his discussion of *code-mixing* in young bilingual children, Genesee (2006:52) writes that evidence that constraints are operative from the outset of two- and multiple-word productions would provide support for the argument that they emerge with the advent of grammatical competence and do not require additional learning. To collect evidence he reviews the research carried out on intra-utterance *code-mixing* by bilingual children in the following language pairs: French and German (Koppe, in press, Meisel, 1994); French and English (Sauve & Genesee, 2000, Paradis et al., 2000); English and Norwegian (Lanza, 1997); English and Estonian (Vihman, 1998), and Inuktitut and English (Allen et al., 2001). The results of these studies support Genesee's conclusion that child bilingual code-mixing *is* indeed grammatically constrained and that these constraints are essentially the same as those that describe adult *code-mixing*. Moreover, he says, they appear to be operative as soon as children begin to combine words into single utterances and, by implication, emerge along with grammatical competence (Genesee, 2006:53).

Recently, research carried out by the Wuppertal Bilingualism Group (WuBIG) at Bergische Universität in Germany examined bilingual children's use of mixed noun phrases (DPs), i.e the use of a modifier (determiners and/or attributive adjectives) in one language and a noun in the other language. Examining spontaneous longitudinal data of mixed utterances of 18 bilingual children acquiring German and either French, Spanish or Italian, they saw that in most cases both types of modifiers were seen to agree with the head noun in switched DPs (Eichler et al, 2013). This was seen as evidence that the children were respecting the grammars of the two

languages, thereby lending support to the theory put foward by MacSwan that no third grammar is needed to explain code-switching.

Another recent study sought to determine whether structural similarity at the level of the noun phrase (NP) across pairs of languages served to faciliate code-switching. Comparing elicited mixed utterances from German-English (G-E) bilingual children to those produced by German-Russian (G-R) children, Endesfelder-Quick (2013) discovered the production of significantly more mixed NPs by the former. She concluded that this was due to the overlap in form and function of NPs which exists in German and English as opposed to the lack of overlap occurring in German and Russian. This study highlights how typological differences can affect the occurrence and frequency of this type of mixing.

Both of the above-mentioned studies were reported on at a conference organised by WuBIG (see above) of which the main aim was to discuss current proposals on code-switching in early bilinguals[4]. Bringing together experts in the field of code-switching (including theoretical rivals Jake (Myers-Scotton's colleague) and MacSwan), eighteen presentations were given over an intense three days. Divided into three thematic blocks, papers reported on research in CS (i) at the clause level, (ii) at the NP level and (iii) on pragmatic and psychological aspects of CS. It was interesting to note that of the 18 presentations, there were only 7 which specifically focused on CS in bilingual children. The remainder (including those given by Jake, MacSwan, Toribio & Bullock, Auer) involved research on teenage or adult bilingual subjects. I would like to posit that despite the WuBIG's desire to focus purely on CS in child bilinguals, their selection of speakers reflects the fact that current code-switching research still prioritises the study of adults as opposed to children. One of the reasons for this, I believe, is the methodological challenges faced by those working with children: certain methods, especially those from within the experimental paradigm (see 2.2.2), would be difficult, if not impossible, to implement with child subjects. Gathering suitable data for analysis, therefore, is problematic and although naturalistic spoken data might provide the best sources for investigating CS (both in child and adult speech), the compilation of such corpora presents its own problems, as will be seen in in the discussion in 2.2.3.

[4]'Code-switching in the bilingual child: within and across the clause' held at Bergische Universitat, Wuppertal, Germany from 18th to 20th April 2013.

Although the focus of the conference mentioned above was more on grammatical issues surrounding CS, psychological and pragmatic aspects were briefly touched upon: Hernandez reported on 'Neural and pyschological bases of switching in bilingual children' (see 2.2.1 for more details) while Hager et al, examined 'Influencing factors on code-switching in a cross-sectional study with German-Romance (…) bilingual children' (2013). In the latter study the researchers looked at the influence of factors such as language dominance, the language of the community, the language of the context and family language policy. In the next section these more sociolinguistic and pragmatic aspects which underly the motivations behind code-switching in bilingual children will be expanded on.

Just as with adults, children may code-switch for a variety of reasons related to cognitive, communicative and social competence. Genesee says that on a cognitive level, a child may code-switch in order to fill a lexical gap (the gap-filling hypothesis). As lexical knowledge in each language will not be the same, a child may draw on resources of the other language to fill the gap of a word for which s/he has no translation equivalent (2006:53). It may also be due to the lack of an appropriate word in the target language (ibid:54). Communicative, or pragmatic, explanations for child CS are the same as for adult CS: to emphasize what they are saying, to quote speakers, to protest, to narrate etc. As with adults, for children one language may have more affective load than the other and be used to express emotion (ibid:55). It is unquestionable that communicative competence (the ability to use languages appropriately and effectively with others) cannot be learnt in social isolation and it is perhaps within the realm of *language socialization* that we can uncover the reasons for a particular child's CS patterns. Due to the importance of the contribution of the field of language socialization to the present study, the next section will be devoted to defining, discussing and detailing aspects of this field of research.

2.1.3.1 Language socialization.

In the introduction to an entire volume about Language Socialization (hereafter LS) (Duff & Hornberger, 2008) Duff describes the process of LS as being 'how children and other novices become both communicatively and culturally competent within their homes, schools, and other discourse communities.' (2008: xiii). This may imply that at some stage an individual's LS is complete. However, it is, in fact, an ongoing process which continues throughout a person's life and is characterized by bi- or

multidirectionality where both novices and experienced members of a community 'are being socialized by mutual engagement in language/literacy practices (...)'(ibid:xv).

Language socialization emerged as a research field following the coming together of two researchers who carried out two separate major longitudinal studies in non-western societies. While Schieffelin studied child-caregiver interactions among the Kaluli in Papua New Guinea between 1975-1977 (Schieffelin, 1985), Ochs focussed on child-caregiver interactions of Samoans between 1977-1979 (Ochs, 1985). At the end of their fieldwork, Schieffelin and Ochs came together in 1984 and fomulated a proposal that presented linguistic and sociocultural development as intersecting processes, stating that:

> '(...) the process of acquiring language is embedded in and constitutive of the process of becoming socialized to be a competent member of a social group and that socialization practices and ideologies impact language acquisition in concert with neurodevelopmental influences.'(Ochs & Schieffelin, 2008:5)

A child is born into 'a lifeworld saturated with social and cultural forces, predilections, symbols, ideologies, and practices that structure language production and comprehension over developmental time.'(ibid:5). Most LS is implicit, with children inferring and appropriating indexical meanings over time and through routine participation in language practices (ibid:9).

The main tenet underlying LS, therefore, is this bi-directional interplay between language and socialization, referred to most frequently in the literature as simply 'socialization *through* language and socialization *into* language.'(ibid:9).

One of the major achievements of research in this field has been to reveal cross-cultural differences in patterns of LS. This can only be effectively done by comparing studies that have chosen as their focus the same kinds of interaction settings. A popular interaction setting is that of mealtimes, the resulting family discourse proving to be, according to Blum-Kulka, 'a major site for the negotiation of linguistic, cognitive, cultural, social, political, and emotive concerns (...)'(Blum-Kulka, 2008:90). In her discussion of dinnertime family discourse, Blum-Kulka cites three important large longitudinal studies which focus on mealtime interactions: Berko Gleason's (1975) study; Snow's (1991) study of 81 working-class families; and Ochs and colleagues' study of 12 American families (1992, 2006). What these studies reveal is how rich dinnertime interactions can be in offering opportunities for exposure to the multifacted aspects of LS:

> '(...) actively listening to the many voices of co-participants at dinner can contribute to perspective taking on truth and knowledge, expose children to multiple varieties of language, to different registers and languages, as well as to different keyings (such as irony and humor), and cultural preferences of politeness and modes of reasoning. Through their own participation in dinner talk, children gain practice in interpreting nonliteral language uses, learn cultural modes of argumentation and giving accounts, acquire cultural notions of tellability and participate in the co-construction of extended texts in various genres.'(Blum-Kulka, 2008:97).

A major part of the recordings in the LOBILL corpus are of mealtime conversations and, as will be seen in later Chapters, they prove to be extremely fruitful in terms of seeing bilingual LS in action, in particular how the children are socialized into code-switching practices.

The link between language socialization in the home and the maintenance of a minority language is another area of research which is important to mention here. Already it is often the case that the mother and/or father are the primary (if not sole) 'transmitters' of the minority language (Morris & Jones, 2008:131) but this situation is becoming further threatened by modern family lifestyles where both parents may work and children spend more time with other carers who may not speak the minority language. These changes will clearly have an effect on family language socialization practices and consequently on code-switching use among family members of bilingual families.

In Wales, a study supported by the Welsh Assembly Government sought to raise awareness among parents of the benefits of bilingualism and of maintaining a minority language. As the study revealed, an important factor influencing language use among family members was the perception of the minority language in terms of economic and social value. Other factors reported on by Morris & Jones included the following:

1. Time spent and interactional practices with minority language speaking parent (i.e what do parents speak to each other/in presence of others).
2. Involvement of grandparents and extended family.
3. The role of older siblings – what they contribute to the youngest sibling's LS.
4. Language background, language values, and language practices of parents and their extended family.
5. Parental language values and power relations, such as who makes decisions about language policy.

(Morris & Jones, 2008:134-7)

All of these factors affecting language use at home will be considered when analysing the code-switching patterns of the informants in the present study.

The role of media as a socializing agent (ibid:138) also needs to be taken into account as activities such as TV, video and DVD watching and story reading have been shown to contribute to minority language socialization. Indeed in the situation where there is no access to the minority language in the community, these often represent the only sources of exposure to the language other than that provided by the parent(s).

Such restricted access to the home, or heritage, language may result not only in a limited command of grammatical and lexical forms but also in an ability to adapt speech according to the requirements of the setting and the audience. As He points out (ibid, 2008:203), ‘‘(...) to know a heritage language means not merely to command the lexico-grammatical forms in both speech and writing, but also to understand or embrace a set of norms, preferences, and expectations relating linguistic structures to context.’ It is pertinent to ask here what effect a drastic change in linguistic and social setting (such as migration) would have on a bilingual child whose primary language socialization was still in progress and whose second language socialization was necessarily being conducted mainly by one care-giver. With a reversal of language exposure and setting, one could predict that given time, the child would become fully socialized into the second language, becoming a competent user, both linguistically and socially. But what about the child’s ‘primary’ language socialization process? No longer exposed to the full array of sociolinguistic settings in her first language, will language attrition set in? What effect will this potential loss of linguistic and social competence in one language have on the child’s CS practices? These are questions which will be investigated in the LOBILL corpus and it is hoped that the answers found will contribute to the little-studied area of the relationship between language shift/attrition and CS in children.

From the discussion above it is evident that language socialization is a key notion for researchers who recognise that differences between bilingual individuals’ linguistic output cannot be explained in purely grammatical terms, that is as a result of typological differences between languages. Equally, however, a purely sociolinguistic approach will not bring us closer to understanding the grammatical

structure of CS patterns found in different language pairs. For researchers who wish to avoid following an exclusively grammatical approach (and aligning, or not, with either Myers-Scotton or MacSwan's models) or an exclusively sociolinguisic approach, there has emerged what could be considered a third 'approach', or rather 'movement' as discussed by Gardner-Chloros (2009). She refers to this 'approach' as the 'Fuzzy school' which places the complex nature of CS at the centre rather than at the periphery (2009:167). Those who 'sympathise' with the 'Fuzzy school' are researchers who recognise the inablity of formal models to explain all types of CS and through their work demonstrate the need to steer clear of bandwagons and approach CS from a more holistic angle. In the next section the key tenets perveying this more holistic approach will be expanded upon.

2.1.4 A holistic approach to code-switching

Having discussed both grammatical and sociolinguistic approaches to CS and examined the models on offer, it has become evident that for a better understanding of the complex nature of CS, especially that occurring in the speech of children, it would be futile to follow one approach exclusively and ignore the insights gained by the other. Both grammatical and sociolinguistic factors have a contributing role to play in determining the outcome of the contact of two language varieties and therefore an eclectic approach would be more enlightening. Furthermore, by comparing CS across different communities and different language combinations, the relative role of linguistic and sociolinguistic factors would have a greater chance of being revealed.

However, in such a holistic approach, it is also crucial to analyse the variation found between speakers in the same community. This could reveal how far idiolectal factors contribute to the different patterns found between individuals, an area of CS research which has so far been neglected according to Li Wei (2002:169). Gardner-Chloros agrees that considering idiolectal competence is key to shedding light on the variation of CS patterns (2009:165). For children, the combination of parental input, sibling language practices and school and community input is likely to result in the development of highly personal and individualistic patterns of CS.

This does not mean, however, that similar patterns cannot be found across speakers within a community or indeed across similar communities in different bilingual language contexts. Research has shown how typological factors can

influence certain aspects of CS. However, rather than talking about this influence in terms of 'constraints', Clyne proposed that we should look at aspects of different language combinations as being 'facilitators' of CS (Clyne, 1987). Gardner-Chloros echoes this proposal when she says that different pairings provide different opportunities and difficulties at a linguistic level, in particular at a syntactic level (2009:166). With regards specifically to CS in children, Vihman found that the qualities of the languages themselves may play a role in CS patterning (Vihman, 1985). For example, English function words may be simpler and more salient than in the other language. By making comparisons between children learning more or less distantly related language pairs, researchers would be able to test the effect of the linguistic factors proper on CS (Gardner-Chloros, 2009:143).

From the discussion above it is clear that a holistic approach to the study of CS involves drawing on knowledge gleaned from examining the role of structural, sociolinguistic and idiolectal factors in an attempt to determine how, and to what extent, they each affect the patterns found in code-switched data. One study, for example, found that when faced with grammatical difficulties (due to typological differences within the language pair), some code-switchers employed bilingual compound verbs as a strategy to deal with these structural difficulties (see Edwards & Gardner-Chloros, 2007). If the same strategy were found to be used by speakers of other language pairs with similar structural challenges, one could postulate the possibility of a universal strategy at play. Thus we have a combination of typological (structural) factors and individuals' idiolectal competence giving rise to innovative linguistic phenomena which cannot be explained in purely grammatical or sociolinguistic terms.

Recognising this interplay between the different factors affecting CS patterns, Muysken sought to propose correlations between types of CS and types of sociolinguistic contexts (Muysken, 2000). Although the typology of code-mixing he proposes does not take into account the effect of idiolectal factors on CS, his descriptive framework is a very useful starting point for researchers who wish to broadly situate their code-switched data from a more holistic angle. As Muysken's typology will be useful when it comes to describing the data in the present study, the following section will present an overview of his framework.

2.1.4.1 A typology of *code-mixing* (Muysken)

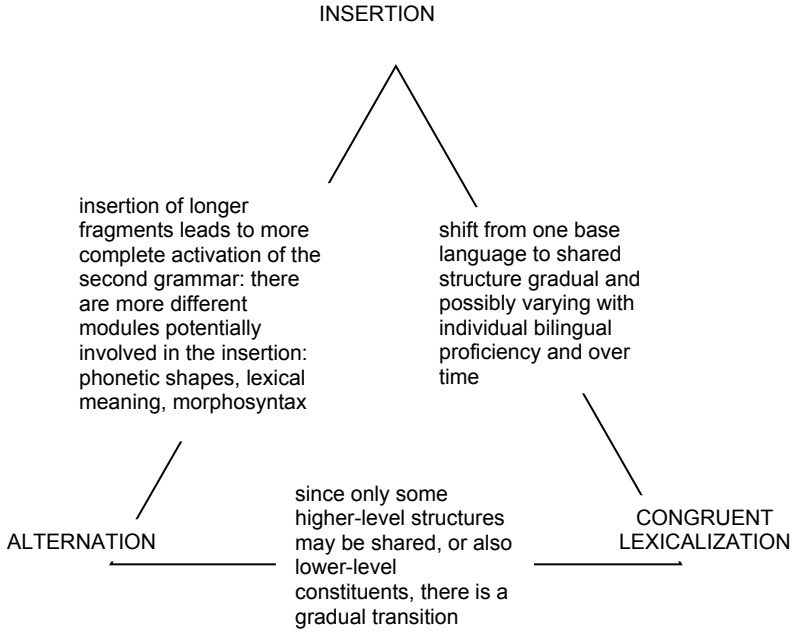
As discussed in 1.2.6, Muysken uses *code-mixing* rather than 'code-switching' as an all-encompassing term, reserving the latter for a certain type of *code-mixing.* In the following discussion the author's use of these terms will be kept in italics so as to remind readers of how they differ from the working definition used in this study (see 1.2.6).

Muysken states that the variation in *code-mixing* patterns encountered in bilingual data is mainly due to the fact that there are three different basic processes at work (2000:3):

- **insertion** of material (lexical items or entire constituents) from one language into a structure from the other language.
- **alternation** between structures from languages.
- **congruent lexicalization** of material from different lexical inventories into a shared grammatical structure.

He places these three types of *code-mixing* at the points of a triangle and illustrates that the differences between them are 'gradual rather than absolute' (ibid:9):

Figure 2. Schematic representation of the three main styles of *code-mixing* and transitions between them (Muysken, 2000)



He goes on to describe the 'sociolinguistic embedding' (ibid:8-9) of these patterns, broadly associating each one with certain types of bilingual communities or settings. His description has been summarized in the table below for clearer visualization:

Table 1. Patterns of *code-mixing* and the sociolinguistic settings in which they frequently occur (adapted from Muysken, 2000)

Pattern of code-mixing	**Types of community**	**Use of languages**
Insertion	Colonial settings, recent migrant communities	Asymmetry in proficiency in two languages
Alternation	Stable bilingual communties	Tradition of language separation
Congruent lexicalization	Second generation migrant groups, dialect/standard and post-creole continua	No tradition of overt language separation

It is important to highlight that Muysken's classification is based on societal bilingualism, involving whole community language practices. In the present study the situation is quite different as it is only within the family unit that bilingual language practices occur: when in Brazil the parents are the only daily source of the 'minority' language (English) and in England the situation is reversed, contact with Portuguese being restricted to parental use in the family home. However, as will be seen in the results section, it appears that a parallel can be drawn between the language processes which occur in migrant communitites (such as a language dominance shift) and those occurring in this particular study. Furthermore, preliminary analyses of the LOBILL Corpus revealed that much of the data examined can be located towards the top of Muysken's triangle, the predominant pattern of *code-mixing* being of the insertional type. However, only after analysis of all the data in the corpus will the predominance of this pattern be confirmed.

Although Muysken's framework will be useful for making the link between the broad sociolinguistic context of this study and the grammatical features of code-switching commonly associated with the setting, it cannot possibly account for the extent to which the code-switching patterns of an individual are influenced by their immediate linguistic input or their underlying linguistic competence, which in the case of a child is still developing. All of these factors will be crucial to take into account when examining the code-switching practices of the informants in the LOBILL Corpus. It is clear that only a holistic approach will make it possible to attempt to

determine the relative role of the various factors mentioned above (typological/grammatical factors, sociolinguistic factors, parental input, idiolectal competence). This type of approach will necessarily require drawing on different methodologies and types of analysis (grammatical, pragmatic, ethnographic, etc) which are typically used in research in CS. It is these methodological aspects that will now be discussed in the following section.

2.2 Methodology in code-switching research

One of the aims of the current study is to present innovative ways of investigating CS in bilingual data - a case will be made for how Corpus Linguistics methodology can provide efficient, effective and novel ways to investigate CS from both a quantitative and qualitative perspective. In order to highlight the advantages of such a methodological approach, it is first necessary to provide a brief overview of some other methods which have been used in CS research. Before looking at experimental research methods, I will first mention those which examine how the brain processes bilingual speech.

2.2.1 Neurocognitive methods

Neurocognitive methods can offer visual insights into the effects CS has on processing. These include hemodynamic techniques such as PET (Positron Emission Tomography) and MRI (Magnetic Resonance Imaging) and electrophysiological techniques such as ERPs (Event-related brain potentials) and MEG (Magnetoencephalography). Participants normally read or listen while they undergo the examination process and measurements show the effects of CS on processing. Recent laboratory work undertaken by Hernandez and colleagues (2013), looked at the correlation between age and the ability to switch between languages. They found that there is developmental improvement in bilingual children's ability to switch (measured in terms of size of switching costs) which is fundamentally related to the development of a more general control or executive function system. The fact that some studies have shown that bilinguals have more developed executive control abilities than monolinguals suggest that 'executive function may be crucial in second language learning during childhood with benefits that extend into adulthood'(2013:15). For readers interested in the latest theory

proposing to account for the improved executive function of bilinguals, one can refer to Stocco et al (2014).

2.2.2 Experimental techniques

In the field of CS research, controlled experimental tasks are most useful for investigating the mechanism of switching itself. Typically, such studies aim to throw light on aspects such as language selection, control, attention, and switching costs, the structure of and access to the bilingual lexicon, and bilingual memory (Gullberg et al, 2009:21).

Single-word level tasks most often involve word recognition and lexical decisions, based on externally induced switching. For example, a naming task may require the speaker to name as many words as possible in a given time window using a particular language or a particular mode; for instance, to name only in one language, to switch language for every word, or to give a translation equivalent for every word offered. For most experimental tasks, results are measured in terms of response/reaction times, and accuracy or error scores.

Sentence-level tasks involve techniques which draw on internally generated switches. In their chapter on research techniques in CS, Gullberg and colleagues mention a total of eleven types of tasks, listed below with a brief explanation for each one:

1. **Grammaticality or acceptability judgement tasks**: traditionally written off-line techniques, there are now auditory versions, participants being asked if the sentence sounds like something they have heard.
2. **Content judgements**: these involve comprehension questions and true/false judgements.
3. **Sentence matching**: participants see 2 sentences on a screen, one slightly after the other and have to press a button indicating if it is the same or not. Response times are measured. In general grammatical sentences are responded to faster than ungrammatical sentences.
4. **Silent reading**: eye-tracking techniques are used to measure reading time. It is generally found that the longer the reading time is the more difficult the processing is.
5. **Auditory moving window**: participants listen to sentences gradually and press a button to receive the next segment. The time for each segment is recorded.
6. **Reading aloud**: participants read texts aloud and speed and influence of switching is measured.
7. **Free speech in "code-switch mode"**: participants are required to speak freely but required to code-switch.
8. **Sentence repetition:** speakers are asked to repeat back sentences as accurately as possible. The rationale is that if the sentence exceeds short-term memory, the listener will pass it through their own grammar before repeating it and this may lead to changes if an element is not part of their grammar.

9. **Sentence completion**: participants fill in blanks as quickly and accurately as possible. This can investigate preferred switch locations.
10. **Sentence recall (priming):** a participant is asked to read and memorize a sentence. Then s/he reads the prime sentence which has a different structure or different language. A distractor question may then be asked before the speaker is finally asked to recall the original sentence.
11. **Confederate scripting**: two participants take turns to describe a picture, one being instructed to use a particular lexical or syntactic construction. The extent to which the real participant used the same construction is measured.

(Gullberg at al, 2009:31-35).

Despite the variety of techniques developed within the experimental paradigm for investigating CS, Gullberg et al recognise a basic flaw with this type of methodology: how do you induce, manipulate and replicate natural CS without compromising the phenomenon itself? (ibid:21). I would also like to add that another major disadvantage of such experiments is that the majority are unsuitable for use with children, not least because in most cases they require subjects to be proficient readers.

Despite such drawbacks, however, experimental techniques have contributed to advances in our understanding of the processes underlying code-switching, especially in adults.

2.2.3 Naturalistic data and early corpus methodology

The use of naturalistic data to research bilingual children's language acquisition and development, including code-switching, dates back to the early twentieth century. In 1913, Jules Ronjat, a French linguist married to a German, published a detailed report of the bilingual language development of his son Louis from birth to 4;10 (Ronjat, 1913). Through field notes, recordings and subsequent transcriptions, Ronjat was able to analyse the data and concluded that a child growing up bilingually suffered from no adverse cognitive effects. In his report he does recommend, however, that parents would be advised to follow the *une personne; une langue* (one person; one language) principle which his fellow linguist, Maurice Grammont, first coined in his 1902 publication *Observations sur le langage des enfants* (Observations on child language)(see Barron-Hauwaert, 2004). According to Grammont's formula, by each parent speaking one language to the child, the chances of 'normal' acquisition of both languages are greatly improved: language mixing (or code-switching) was seen as something which needed to be avoided. Another important longitudinal naturalistic case study was that carried out by Werner

Leopold (of German nationality), who, between 1939-1949, followed the language development (in German and English) of his two daughters, Hildegard and Karla (Leopold, 1939-1949, 1970). His data also led him to arrive at the same conclusion as Ronjat about the importance of adhering to a *one-person-one-language* principle in order to foster 'normal' monolingual acquisition of both languages and minimise language mixing.

Although some of the conclusions arrived at by the above two studies may now be disputed by many researchers, it is their importance in terms of methodology which is the focus of this section - both studies showed how insightful the collection and analysis of naturalistic data can be. Even more so if the observers are in the priviledged position of being both researchers and parents of the children under observation. Ronjat's and Leopold's studies marked the beginning of a tradition of researcher/parent bilingual case studies, more modern ones of which include the following: the works of Marilyn Vihman (Estonian/English), Margaret Deuchar (Spanish/English), Annick de Houwer (Dutch/English), Virginia Yip and Stephen Matthews (Cantonese/English), Phillip Carr (French/English), and Philip and Elizabeth Prinz (ASL/English)[5]. As far as notable trilingual case studies are concerned, one should mention the parent/reseachers Jean-Marc Dewaele (French, Dutch, English) and Madalena Cruz-Ferreira (Portuguese, Swedish, English)[6].

Until relatively recently, the traditional method for collecting naturalistic data by researchers such as those mentioned above was via field notes and audio and/or video recorders. With the advent of new digital technologies the recording and storing of participants' language behaviour has become easier, much more unobtrusive and often a multimodal experience, capable of capturing speech, facial expressions and gestures. These advances in technology have made it possible for researchers to amass large quantities of data which have the potential to provide new insights in many areas of linguistic enquiry. It is clear that traditional manual analyses of such data becomes less and less feasible as the size of the corpus increases and this has necessarily led to the development of specialized computer software specifically designed to assist researchers in being able to effectively exploit the data they collect. Such methodology belongs to the relatively new field of Corpus Linguistics

[5] Selected papers include the following: Vihman (1985, 1998); Deuchar & Clark (1996) and Deuchar & Quay (2000); de Hower (1990, 2009); Yip & Matthews (2000, 2007); Carr (2007) and Brulard & Carr (2003); Prinz et al (1985) and Prinz & Prinz (1979).

[6] See Dewaele (2001, 2007) and Cruz-Ferreira (1999, 2006, 2010).

(CL) which, especially over the last 20 years, has seen an explosion in terms of both the creation of new corpora (written and oral) and the number of studies dedicated to the uncovering of previously undetectable patterns in this new language data. In the following section I will discuss why, despite such advances in the electronic investigation of language, there appears to be a notable lack of progress when it comes to using CL techniques to investigate code-switching in spoken corpora.

2.2.4 Corpus Linguistics and code-switching

Poplack's quantitative analysis of surface configuations of code-switches in a corpus of Spanish/English spontaneous data (Poplack, 1980) represented a significant contribution to the field in terms of methodology as it demonstrated the importance of using empirical data for CS analysis as opposed to a much reduced selection of utterances, often consisting of made-up examples, as those used in grammaticality judgment tasks. With 66 hours of recordings of the spontaneous speech of 20 participants, Poplack was able to subsequently analyse 1,835 instances of naturally occurring code-switches. Although at the time of her study, manual extraction and analysis of the data would perhaps have been the only method available, we would now expect to be able to exploit such a corpus using modern automated techniques. As the current study will demonstrate, the use of software has several advantages over manual methods, not least the instant extraction of key items and the potential to reveal patterns in data which may otherwise go unnoticed.

However, the compilation of a corpus of spoken language which can be analysed using modern software is no mean feat. Although researchers now have tools available to them to create all types of written corpora, some of which involve the relatively simple process of 'web-crawling' (see Biemann et al, 2013 for a comprehensive look at the construction of high quality web-based corpora), the creation of oral corpora is much more challenging. In a recent article in which Travis and Cacoullos (2013) write about the compilation of their bilingual corpus, they highlight the time expended on transcription saying that approximately 50 hours was spent on transcribing every hour of recorded data. They add that even more time was spent on multi-party conversations. This raises a very important question which reflects a criticism often sited by researchers working within experimental paradigms - can one really justify the time expended on the transcription of a corpus that might only ever be examined by the researcher or the team that built it (Gullberg et al,

2009: 23)? Indeed these authors go on to state that ‘virtually none of the bilingual corpora on which the CS studies are based are publicly available. It is therefore not possible to study the same materials in order to test the conclusions reached or explore other interpretations’(ibid: 23).

It is understandable that ethical issues may often prevent the release of spoken corpora into the public domain (Adolphs & Carter, 2013:10-11) and although it is possible to gain access to some bilingual corpora via written permission, even then the outside researcher is faced with two additional methodological problems: firstly, project internal transcription conventions may make it difficult to access (understand) the data; secondly, the specific format of the transcriptions is likely to mean that project external software cannot be used to examine the data.

Faced with such challenges, one might be inclined to agree with Gullberg et al when they argue that it would be more productive to test hypotheses, arising from empirical studies, using experimental research methodology (ibid:22-26). However, as stated in my introduction (1.3), there is still much to be discovered about bilingual phenomena and it is only by examining different sets of naturalistic spoken data that CS patterns and trends will become more apparent. But how to proceed in the face of so many challenges? When one learns that a project, coordinated by Brian MacWhinney and launched in 1981, already appeared to offer solutions to these methodological problems, one can only suppose that Gullberg was still unaware of such a project even after 18 years of its existence (his criticisms were made in 2009). Indeed, I would add that, unfortunately, this 'lack of awareness' is widespread and is most likely the result of the (political) nature of each university's own research agenda where innovation is a priority, even if this means 'expending large amounts of time or resources or having to start from scratch, each time a spoken corpus (…) is required' (Adolphs & Carter, 2013:12).

Recognising the need for greater collaboration among researchers wishing to study spoken language, MacWhinney and colleagues (Dan Slobin, Catherine Snow, Willem Levelt and Susan Ervin-Tripp) conceived a multilingual project they named CHILDES (Child Language Data Exchange System) which aimed to propose solutions to the challenges discussed above. As the project they developed is of fundamental importance to my study, full details will be provided in the following section.

2.2.4.1 CHILDES (Child Language Data Exchange System)

As mentioned above, CHILDES is a multilingual project originally conceived in 1981, by Brian MacWhinney and colleagues in the United States (MacWhinney & Snow, 1985, 1990). The original objective of the project was to establish a standardized and systemized system of transcription that would permit the exchange of data stored in electronic corpora between researchers all over the world, thereby making progress in the study of language acquisition (first and second) possible (MacWhinney, 1991).

Freely available over the Internet, CHILDES is constituted of the following three components:

1. The CHILDES database (MacWhinney, 2014a), which contains corpora from various sources, contributions from more than 100 researchers who work in the area of language acquisition. There are data from various languages, and, being an open data base, any researcher can contribute his data, as long as they are transcribed in the CHAT format (see below).
2. CHAT (Codes for the Human Analysis of Transcripts)(MacWhinney, 2014b), which consists of a system of transcription which has the objective of making the coding of corpora uniform in order to permit different types of analyses through the use of electronic tools.
3. CLAN (Computerized Language Analysis)(MacWhinney, 2014c) which is a programme specifically developed to analyse data transcribed in CHAT format.

CHAT is a very clear and complete transcription system which is easy to apply to data and easy to read. Apart from the transcription itself, the CHAT system includes a series of conventions for the process of coding various linguistic phenomena. As long as the corpus is transcribed following the standards suggested in the CHILDES manual, the programme of analysis, CLAN, can be used to exploit it. The possibilities of analyses increase with the insertion of codes, the choice of which will vary according to the objective of the researcher. In Chapter 3, more detail will be given as to how CHAT was used to transcribe the corpus investigated in this study.

2.2.4.1.1 Bilingual data in CHILDES

The bilingual section of the CHILDES database currently offers 21 bilingual corpora, and 1 trilingual corpus, which can be downloaded and analysed using the CLAN tools. The different combinations of language pairs are shown in Table 2. Note that

the language in the first column does not imply that this is the more dominant language of the two or is the 'first' language of the bilingual speakers involved. The numbers in brackets indicate how many of that particular language pair can be found, any absence of number meaning the existence of only one corpus.

Table 2. Combinations of languages found in the bilingual corpora available through CHILDES (adapted from MacWhinney, 2014a)

English +	Chinese (2), Dutch, French (2), Polish, Russian, Spanish (4), Persian & Hungarian (i.e. trilingual)
Dutch +	Arabic, Turkish, Italian
Danish +	Japanese
French +	Chinese
Italian +	Austrian German
Portuguese +	Swedish
Spanish +	Catalan (2)
Russian +	German

As is evident from the table, over half (12 out of 22) of the bilingual corpora have English as one of the languages of the pair. However, despite there being a Portuguese/Swedish corpus, there is currently no Portuguese/English corpus of spoken language in existence within CHILDES. My corpus would be the first of its type to be made freely available to the linguistic research community. In addition to its originality in terms of combination of languages, the corpus in this study also aims to offer a significant contribution in terms of improvements to how bilingual data is currently coded in bilingual corpora. This is discussed in the following section.

2.2.4.1.2 Language coding in CHILDES' bilingual data

On pages 93-4 of the CHAT manual (2014b) we find suggestions for the coding of code-switched material. Of these, the one which would allow for most comprehensive retrieval and study is where each word is coded individually. For example, 'ball' could be transcribed as 'ball@e' (where 'e' is 'English') and 'bola' would be transcribed as 'bola@p' ('p' representing 'Portuguese'). Understandably this method of language coding would be very time-consuming and it is perhaps in order to avoid discouraging potential contributors that this, or any type of language coding is not made obligatory under the CHAT system. Unfortunately, however, this means that apart from some of the bilingual corpora lacking any language coding whatsoever, others have chosen to use one of the other coding suggestions. These differences and inconsistencies in

coding across (and even within) the bilingual corpora has implications in terms of analyses: it impossibilizes, or at best limits, the replication of certain CLAN analyses which would allow for the comparative analysis of various aspects of code-switching across the different language data.

Recognizing the need for consistency in terms of coding languages transcribed according to the CHAT system, a group of researchers (including Mark Sebba, Penelope Gardner-Chloros and Melissa Moyer) came together to form the LIPPS Group (Language Interaction in Plurilingual and Plurilectal Speakers)(see http://ling.lancs.ac.uk/staff/mark/lipps/lipps) and set up an offshoot of CHILDES that was named LIDES (Language Interaction Data Exchange System). Based entirely on the CHAT system, in reality the only improvement the LIDES transcription system (Gardner-Chloros et al, 2000) offers is the *obligatory* use of language coding for each word, as illustrated above. Although in order to facilitate comparisons across different language pairs, recommendations are also made for the insertion of glosses or translations into English of utterances in other languages, this is also originally suggested in the CHAT manual. One justification for the creation of a separate data bank to that in CHILDES might be that the bilingual corpora in LIDES contain the speech of older bilinguals, that is teenagers and adults, as opposed to children. However, they could easily have been included in the BilingBank Database which is where non-child bilingual data (transcribed in CHAT) is currently stored under the larger umbrella database called Talkbank, (see Talkbank.org) coordinated by Brian MacWhinney.

Ideally, all the bilingual corpora transcribed in CHAT format would be kept in one repository and use the same language coding. This would enable a multitude of cross-corpora analyses of code-switching. As it stands the main impediment to this occurring is the time it takes to transcribe a bilingual corpus: coding every single word is monotonous and time-consuming (Gardner-Chloros et al, 2000:139). Moreover, with this method of coding, the transcripts become more difficult to read, making qualitative analysis of the transcripts more challenging. What is clearly needed is a method of language coding which would be less time-consuming and which would increase readability of transcripts while still allowing for maximum automated exploitation of the bilingual data.

Apart from proposing such an improved method of language coding in this study (see Chapter 3) I also seek to address what I believe to be an important

oversight on the part of the LIPPS researchers in relation to their coding recommendations for CS data. In their manual there is no insistence on the insertion of addressee codes (codes which identify to whom each utterance is directed). If one considers that the addressee is a crucial variable when analysing both the motivations underlying CS and its grammatical nature, it would appear that the LIPPS group are unneccessarily limiting the types of analyses which can be carried out their data. This will become evident in the discussion of my results in Chapters 4 to 7.

2.2.4.1.3 Language analyses of bilingual CHAT data

Apart from outlining improvements to the current CHAT/LIDES coding system, another significant contribution which this study aims to offer is related to the performing of analyses on bilingual corpora using CLAN. Although the second goal of the LIPPS group was to put forward specific ways of analysing bilingual corpora, despite reporting on some results (Gardner-Chloros, 2009), it appears that little has been done to share common methods of analyses using the CLAN tools. In fact, it was my frustration at not finding details about *how* a bilingual corpus could be exploited, that prompted my decision to be extremely transparent in my methodology, detailing my analyses to such an extent that replicability would be possible on my own corpus or on any other corpus transcribed in the same way. While the specific CLAN commands which I used are discussed in Chapter 3, it is in Chapters 4 to 7, when I report on my results, that the construction of each command line is detailed.

As is the case with most CL studies, frequency analyses are used to uncover most of the patterns present in my corpus data. However, as will be shown over the next two chapters, it is my use of two other types of quantitative measures, that of word length and vocabulary diversity, which perhaps provide the most original contribution of this study in methodological terms. These measures, traditionally used to analyse a monolingual speaker's language development and vocabulary diversity, are applied in such a way as to enable the investigation of the language asymmetry typically found in the code-switched speech of bilinguals. This methodological innovation constitutes one of three original contributions that I propose to make through the current study. Further details of such methodological aspects of my research will be found in the next chapter to which we will presently turn. First, however, it is necessary to summarize the rationale behind my study and state my research questions.

2.3 Rationale and research questions

The literature review in this chapter has served to highlight the need for studies like the one reported on in this dissertation. First of all, it has become evident through the discussion of previous CS research (2.1), that rather than approach this field from a purely grammatical or sociolinguistic angle, what is needed is a holistic approach which aims to investigate the relative roles of typological, sociolinguistic and idiolectal factors in code-switching behaviour. As such, the major questions that I will be asking of my data are those stated below:

Does the code-switching data found in the bilingual LOBILL Corpus (English/Portuguese) provide support for the MFL and 4-M Models? That is, can these models successfully account for the code-switching patterns of the informants?

How are each informant's code-switching practices affected by (i) community-level factors (i.e the language practices of the community) (ii) family language practices and (iii) conversational-level factors (including interaction type and addressees)?

To what extent is each sibling's code-switching practice affected by their (developing) idiolectal competence?

It is important to point out that the addressee coding in the corpus makes it possible to investigate these and several other questions in the corpus data. I am able to examine any combination of speaker(s) and addressee(s) and output data which might provide answers to questions such as the following:

Does the mother always use English when addressing her children?
How much Portuguese do the siblings use with their mother?
Do the siblings address each other in Portuguese and/or English?
When addressing multiple bilingual speakers (i.e both parents), is English or Portuguese, or a mixture of both, preferred by the siblings?
What effect does the presence of a monolingual speaker have on the children's language use?

In order to seek answers to these various questions I need to be able to analyse my data both quantitatively and qualitatively: while frequency analyses will reveal much

about the CS patterns for each speakers, it is only through a more qualitative analysis that it will be possible to shed light on sociolinguistic and pragmatic motivations underlying their use of CS. Rather than using more traditional (manual) ways to analyse the naturalistic data, I propose to show how more modern techniques, drawn from the field of Corpus Linguistics, can be used to provide unique perspectives on the data. Although now commonly used in monolingual linguistic enquiry, studies reporting on the application of such CL methodology to CS data are notably lacking and, as discussed in 2.2.4, this appears to be the result of the lack of freely available, appropriately coded spoken corpora. While CHILDES does offer valuable access to bilingual corpora, improvements to the system of language coding would greatly increase their exploitability, as will be demonstrated in this study.

By detailing the processes of compilation, transcription and coding of the corpus and showing how to perform innovative analyses, it is hoped that the current research will represent not only a significant contribution to the field of code-switching research but also to that of Corpus Linguistics in general. Furthermore, by contributing the corpus itself to CHILDES, this means that original data for the language pair Portuguese/English will be made available to the wider academic community for further linguistic enquiry and for cross-linguistic comparative research. Although details about the corpus will be forthcoming in the next chapter, it is important to highlight here that the longitudinal nature of my corpus (with over three years of data) means that it provides a particularly rich data source: it will allow me to examine how the CS practices of the siblings develop and change over time.

The present study thus offers a three-fold contribution: original corpus data, original methodology and original results. While Chapters 4 to 7 are dedicated to the discussion of the results of the analyses performed on the corpus, it is in the next chapter (3) that I will introduce the corpus itself and describe how it was transcribed and coded in order for said analyses to be carried out.

3. Methodology

In this chapter details about the research design of this study will be given. The first part of the chapter will be taken up with the description of the LOBILL Corpus and how it was constructed. After presenting information regarding the informants and the procedures for data collection, details about the transcription process and the specific coding system developed for the corpus will be given. The second part of the chapter aims to present the methods of analyses used on the data, described in such detail so as to enable replication of these analyses on the LOBILL Corpus or any other corpora transcribed according to CHAT conventions.

3.1 The LOBILL Corpus

The longitudinal data used in this research form part of the LOBILL Corpus which is composed of the spoken language of two bilingual children in their interactions with mono and bilingual interlocutors, in diverse family situations. The name of the corpus is made up of the combination of the first two letters of the informants' surname, LO (from LOnngren), the abbreviation BIL (from BILingual) and L (from Language). This section aims to provide details about the main informants and their sociolinguistic context (3.1.1), to briefly describe the siblings' language experience (3.1.2), to outline the method of data collection (3.1.3), to describe more specific characteristics of the data (3.1.4) and explain the rationale behind the naming of files (3.1.5).

3.1.1 The informants

The main subjects of the corpus are a brother, JAM[7], and his older sister, MEG, who were 3;5 and 5;10 respectively at the beginning of the data collection. Both were born in Fortaleza, Brazil and attended a Brazilian school from the age of 1;6 until they moved to England in 2004 when they were 6;3 and 8;7. Their mother, MOT, is the researcher and is a native speaker of British English and near-native speaker of Portuguese: after studying Portuguese and Spanish at university in England she then lived in Brazil for twelve years before returning to England in 2004. She is married to a Brazilian, PAI, who speaks English fluently (learnt during a year long stay in

[7] In this dissertation the informants are indentified by identity codes that are used in the transcriptions: following CHAT conventions, the code is composed of a combination of three capital letters, which can be based on the speaker's real name, for example, JAM (James) and MEG (Meggie), or their role, for example, MOT (Mother) and PAI (the Portuguese word for 'father'). See section 3.3 for more details.

England). Although these four bilingual family members, JAM, MEG, MOT and PAI, can be considered the main informants of the study in terms of the amount of recorded data, the participation of other (monolingual) speakers in the recordings, such as grandparents and cousins, is a significant feature of the design of the LOBILL Corpus. This significance will be revealed in the discussion of the results.

3.1.2 The siblings' language experience

In terms of language experience it is possible to separate the corpus data into two major phases which correspond to before and after moving to England. As will be seen in the results this change in their social, cultural and linguistic milieu was to have a significant effect on their linguistic practices. Therefore, I will briefly describe the linguistic journey of the siblings (and their parents) which occured within the time frame of this study[8].

3.1.2.1 From birth until the move to England

Before the birth of their children, Portuguese formed the basis on which all daily interaction between MOT and PAI took place, although code-switching did occur. From the birth of MEG in 1995 the family language dynamics changed: while MOT spoke exclusively English to her daughter PAI used Portuguese when addressing MEG. This daily use of English at home led to greater use of English between the parents, mostly in their code-switching practices. This pattern was further consolidated when JAM was born in 1998, MOT continuing to speak English to both siblings while PAI interacted with them mostly in Portuguese. Other daily inputs of English were restricted to television programmes (Cartoon Network and Discovery Channel) and English story books (read by the mother). Occasional visits from English relatives provided another important source of contact with English and most years both children spent short periods on holiday in England with their mother, where they stayed with their English Grandmother (1996, 1998, 2000 and 2003).

Despite the mother's use of English to both children, the interaction between the siblings was predominantly in Portuguese, following the model of interaction experienced with their peers at their Brazilian primary school. While in England on holiday, there was more use of English between the siblings, especially when in the

[8] Although a total of 151 recordings have been made to date, only the first 119 (up until December 2004) were used for analysis purposes. Therefore, the description of the siblings' linguistic experience will not go beyond this period.

presence of English cousins. When the family moved to England in June 2004, JAM was 6;3 and MEG was 8;7. While MEG had been reading and writing in Portuguese for 2 years, JAM had only just learnt to read and write in Portuguese. Although MEG was able to read in English, her written English showed clear influence from Portuguese. JAM was able to read some English but there was no evidence that he was able (or unable) to write English words.

3.1.2.2 After the move to England

Immediately after arriving in England, MEG, JAM and their mother stayed with the children's English grandmother, GRA, and their auntie, BEC. Their father, PAI, was due to arrive two months later, in August. They began primary school three days after arriving and thus both at home and at school they were immersed in English. For the next two months the children's only source of Portuguese were their interactions with each other and telephone calls to their father in Brazil. Their mother continued to interact with them mostly in English. With the arrival of their father at the end of August and a move into a family home of their own, Portuguese again began to play a more prominent role in the siblings' home environment. Parental attitudes towards the use of Portuguese and code-switching were positive, that is, JAM and MEG were not discouraged from using Portuguese whether in monolingual or bilingual mode. Despite their father's continued use of Portuguese at home, it became apparent over the next few months that English was beginning to feature more and more in the siblings' interactions with their parents and with each other. Although the data analysed in this study only covers the time period up to December 2004 (approximately six months after the family's arrival in England), it is possible to predict that Portuguese would struggle to survive in such an English-dominant environment[9].

From the description above it is evident that it would be impossible to place MEG and JAM on a fixed point on the bilingualism continuum. Over time the balance of their two languages has changed, affected mostly by contextual factors which have increased/reduced their exposure to English and Portuguese. As will be seen in the results this change in exposure clearly affected their code-switching practices.

[9] Field notes and recordings carried out in 2007 did indeed show that Portuguese was very rarely used by the siblings when interacting with their parents and with each other.

3.1.3 Data collection procedures

The main procedure for the collection of data was naturalistic observation, following in the tradition of other researchers (as discussed in 2.2.4). With the aim of capturing the bilinguals' linguistic behaviour in diverse contexts, recordings of MEG and JAM were carried out in various situations, examples of which include meal times, playing board games, at the airport while seeing off a visiting relative and on the telephone. Most of the time the informants were fully aware of the presence of the recorder; that is the recorder was never deliberately hidden[10]. With time this meant that the speakers became accustomed to its presence with the recorded interactions appearing to exhibit uninhibited natural conversation[11].

The period of data collection began in August 2001 and finished in December 2004. With a total of 119 recordings carried out over 3 years and 4 months, this led to an average of 3 recordings per month. However, it is important to note that the time interval between recordings was not necessarily regular. For example, there are some months where only one recording was made while in other months, especially those pertaining to holidays in England, more than 7 recordings were carried out. The length of each recording was mostly determined by the nature of the interaction and varied between as little as 3 minutes (a short phone call) to over 50 minutes (a game of 'Guess Who'). In total, the 119 recordings amounted to just over 24 hours of spoken data. Despite the lack of consistency in terms of regularity and length of recordings, the longitudinal nature of the data still afforded a unique investigation into the code-switching practices of the siblings, and their parents, as will be seen in the results section.

With a view to further comparative analyses, a second set of recordings were carried out between February 2007 and October 2009, amounting to a total of approximately 11 hours of spoken data distributed amongst 33 recordings. Thirteen of these involve JAM and MEG talking to their Brazilian relatives over Skype and as such they have the potential to reveal much about the state of the siblings' Portuguese three years on from their arrival in England. However, within the time constraints of this study, it was not possible to transcribe these interactions and

[10] Ethical approval was duly obtained.

[11] A notable exception to this can be found in some of the 'interview' data with MEG where there is evidence that her linguistic behaviour is directly affected by her awareness of the recorder. This is discussed in Chapter 6.

therefore they do not form part of the LOBILL Corpus analysed here and subsequently submitted to the CHILDES data bank.

With regards to the more technical side of data collection, initially (between 2001 and March 2007), the data were collected using a mini-cassette recorder. A digital recorder was then purchased to replace the mini-cassette recorder. This meant that the recordings could be electronically stored, a safer and more convenient method of storage. With data on mini-cassettes going back to 2001 it was then necessary to convert 25 hours' worth of recordings to a digital format. A rather lengthy but necessary process, this was achieved by re-recording each recording from the mini-cassette recorder to the digital recorder via a cable. The first 119 recordings were then transferred to a specific folder in the computer called the LOBILL Corpus and back-up copies were made. All other later recordings were stored in a separate folder.

In terms of transcription, most recordings were transcribed as soon as possible after the event in paper format and extra-linguistic notes were added to aid later analysis. These basic transcriptions were then inputted into the CHAT text editor at a later date and coded (see section 3.2). However, whenever there was direct access to a computer with the CLAN software, it was possible to transcribe directly into the CHAT editor. Transcriptions were checked several times but unfortunately it was not viable to involve any additional transcriber in this verification process. Therefore, any transcription errors are my own. Details about the transcription system used can be found in section 3.2 as can the specific coding chosen to annotate the LOBILL Corpus.

3.1.4 The data

The general nature of the LOBILL Corpus was outlined above in 3.1.1 and 3.1.2. In this section more specific details will be given regarding the different characteristics of the data and how these translate into six variables which allow for a indepth investigation into the different factors which affect the code-switching behaviour of the bilingual siblings. The first three variables are related to the roles of the participants in the corpus, as discussed below.

3.1.4.1 The speakers, interlocutors and third parties

We already know that the bilingual siblings, JAM and MEG, are the two main informants of the study, appearing in each recording either together or separately. While their bilingual mother features in every recording, their bilingual father appears less frequently. Although it is the interactions within this bilingual family unit which provide most of the data for analysis, the recordings in which monolingual speakers also appear allow for the investigation of the effect of a speaker's monolingual status on the code-switching patterns of the siblings.

In total there are 15 monolingual speakers who feature in the LOBILL Corpus: six British speakers of English and nine speakers of Brazilian Portuguese. Of the latter, only two are not native Brazilians: DAN and his son VIN. Although they are Swedish, as they only speak Portuguese with the main informants of the study, they were classified as monolingual Portuguese speakers. The table below shows the three-letter speaker codes chosen for each participant. They are listed in alphabetical order within each language category.

Table 3. Monolingual speakers who feature in the LOBILL Corpus

Language of speaker	**CHAT speaker Code**	**Relationship to JAM and MEG**
English	BEC	Aunt Becky
	GRA	Grandmother (maternal)
	GRD	Grandfather (maternal)
	JAK	Cousin Jake
	MAX	Cousin Max
	WIL	Uncle William
Portuguese	ARL	Arlene, the maid
	AVO	Grandfather (paternal)
	DAN	Friend's father Dan
	JAN	Uncle Janus
	JUL	Cousin Julia
	ROS	Aunt Rosa
	SAR	Cousin Sara
	VIN	Friend Vincent
	VOV	Grandmother (paternal)

The contribution each speaker makes in terms of tokens varies substantially and there are seven participants, GRD, DAN, JAN, JUL, ROS, VIN and VOV who have a zero token count. This is because they only feature as interlocutors over the telephone, their turns not recorded but merely indicated in these particular transcriptions. Although there are three other participants who speak with the siblings

over the telephone (GRA, AVO and SAR) they also feature in face to face recordings with the siblings and therefore their token count is not zero.

One might question the usefulness of analysing transcripts where half the data is missing (i.e. not recorded). However, despite the apparent 'one-sidedness' of these telephone conversations, the siblings' data is still valuable as it can be analysed in terms of the mono/bilingual variable afforded by the status of each interlocutor. In addition, it will be possible to see how JAM and MEG's language use, and code-switching practice, is affected by a medium of communication which is devoid of paralinguistic cues.

Apart from being able to analyse the data in terms of whom is speaking to whom (i.e. the speaker/interlocutor combination), a third variable which can also be investigated in terms of its effect on the siblings' code-switching practices is that of the presence of other participants in the interactions. For example, while playing a card game with their monolingual English cousins, one could examine whether, when addressing each other, MEG and JAM use Portuguese (their normal mode of communication), whether they code-switch or whether they use English in order to accommodate the presence of a monolingual speaker. While a complete list of participants per file can be found in Appendix A, the technicalities of being able to incorporate the variables of speaker, interlocutor and third party presence in my CLAN analyses of the data will be discussed in detail in section 3.2.

3.1.4.2 Interaction types, location and time periods

Apart from the three variables mentioned above which are specifically related to the informants, the heterogenous nature of the LOBILL Corpus also allows for the investigation of the effect of three other variables on the siblings' code-switching practices. These are more contextual in nature, namely the type of interaction involved, the location of the recording and when it took place.

In 3.1.3 four types of interactions were mentioned: meal times, playing board games, at the airport while seeing off a visiting relative and on the telephone. The LOBILL Corpus contains several other types of recordings and for analysis purposes I decided to group them into one of seven interaction types: Meal Time interactions (MT), Telephone Interactions (TI), Playing Games (PG), Free Play activities (FP), Chatting (CH), Literacy Activities (LA), and Interviews (IN). Examples of the types of recordings which fall into each broad category can be seen in the table below:

Table 4. Interaction types which make up the LOBILL Corpus

Interaction type	**Examples**
Meal Times (MT)	breakfast, lunch, dinner
Telephone Interactions (TI)	calls between the siblings (in England) and their father (in Brazil) and other Brazilian relatives, calls between the siblings (in Brazil) and their British grandparents (in England)
Playing Games (PG)	structured games such as board games, card games etc
Free Play (FP)	activities such as playing with bricks and other toys, painting etc
Chatting (CH)	conversations focussing on events removed from the immediate context e.g. talking about a school trip
Literacy Activities (LA)	reading stories out loud, telling stories from pictures
Interviews (IN)	structured ‘chats’ where the mother asks questions to each sibling

This heterogeneity in terms of interaction types is a marked, and perhaps unique, characteristic of the LOBILL Corpus and allows for comparisons to be made across interaction types, as will be seen in the results section.

The effect of geographical location/sociolinguistic environment on code-switching behaviour is a further variable which can be examined in the LOBILL Corpus. Although the majority of the recordings (65%) were carried out in Brazil, a significant number (35%) occurred in England. In Brazil the specific locations included the family home, the family beach house, a friend's house in the mountains and at the airport. In England the interactions took place at the Grandmother's flat, the new family home, the cousins' house and the Grandfather's holiday home. Specific information about both interaction type and location per file can be found in Appendix A. The latter also contains details relating to the third variable of time, such as the month and year of each recording and the age of the siblings. The incorporation of this 'time' variable is briefly discussed below.

Details about the longitudinal nature of the corpus were provided in 3.1.3. With data spanning 3 years and 4 months one might consider simply dividing up the data into different time periods of, for example, three months as this would allow for a developmental perspective of the siblings' code-switching practices. However, such a simplistic division of the corpus would be problematic for two main reasons: firstly, some time periods would include recordings carried out in two different locations (Brazil and England) and secondly, the number of recordings (and therefore the amount of data) would vary between periods. In any case, the division (or not) of the data into different time periods was actually determined by the type of analyses being

performed. Therefore, specific details about how I exploited the longitudinal nature of the corpus will be found in the discussion of those particular analyses where this time variable was included.

Through this more detailed discussion of the data, I have shown how the heterogeneous and longitudinal nature of the corpus lends itself to the investigation of six variables which can affect code-switching practices: those of speaker, addressee, third party presence, interaction type, location and time. It is evident that the exploitation of such variables will lead to much richer interpretations of the data in the corpus. Indeed, it could be said that in order to arrive at correct interpretations of the bilingual siblings' code-switching behaviour, it would be unwise, or even impossible, *not* to consider the effect of any of these variables.

In practical terms, the investigation of these different variables is achieved by combining the simple process of file selection with the construction of specific CLAN command lines, the latter being intrinsically linked to the type of coding used to annotate the corpus data. Before presenting the types of codes used in the LOBILL Corpus, I will first explain the rationale behind the naming of the 119 files which make up the corpus.

3.1.5 Naming the files

In CHAT there is no specified format for naming files and it is up to the researcher-transcriber to decide how they wish their files to be named. Some examples of file names found in the CHILDES data base are the following[12]:

anne03a.cha (from the 'Manchester' British English corpus)
pa003.cha (from the 'Florianopolis' Brazilian Portuguese corpus)
k17.cha (from the 'De Hower' bilingual English-Dutch corpus)
mar22.cha (from the 'Krupa' Bilingual English-Polish corpus)

The file names are typically made up of a letter, or letters, which indicate the name of the child (**Anne; Paulo; Kate; Martin**) and the number of the file (**03a; 003; 17; 22**). The **.cha** indicates the format of the file (as opposed to **.doc** for example).

As one of the aims of the present research is to contribute a corpus to the research community which can be easily analysed by others, anything that will assist

[12] For information about these corpora see the database manual (MacWhinney, 2014a) and click on **3englishbrit.pdf**, **8romance.pdf** and **4biling.pdf**

a researcher in the selection files for analysis purposes was considered useful. It was with this consideration in mind that the particular file name format for the LOBILL Corpus was arrived at, illustrated in the following four examples[13]:

003INenMSEP01.cha

009CHenJJUN02.cha

027PGenJ&MNOV02.cha

060TIptJ&MJUL03.cha

Each header entry begins with a number, **001** being the first recording, **002** the second and so on. This is followed by a two-letter code which indicates the interaction type (above we have **IN** for Interviews, **CH** for Chatting, **PG** for Playing Games and **TI** for Telephone Interaction). Next comes the code of the language which was considered to predominate (in terms of quantity) in that particular interaction, **en** for English and **pt** for Portuguese[14]. Following the two-letter language code, the letters **J, M** or **J&M** were inserted, thereby telling the researcher which child was involved in the interaction, the latter being used when both children were present. The final code indicates the month and year of the recording, especially useful if tracking linguistic phenomena over a time period. The **.cha** extension simply shows that the file is transcribed in CHAT format.

As will be seen in the discussion of the results, by naming the files in this way, certain types of comparative analysis were easier to perform. For example, certain types of interactions involving both, or either, children could be compared longitudinally: those occurring in a three month period in 2001 could be compared with the same period in 2002 or 2003 etc. The pertinent files could be easily selected via the file name alone, making it unnecessary to open up each file, which would be more time-consuming. This also applies to the selection of files according to the dominant language of the interaction. Already included in the file name, it is not necessary to open individual files in the corpus to find out the main language of the interaction - by the file name alone it is possible to restrict the selection for analysis

[13] The complete list of file names can be found in Appendix A.

[14] Updates to CLAN mean that three-letter language codes (**eng** and **por**) are now required in the **@Languages** header found at the top of each file transcript (see Appendix B1.2.2). However, the use of **en** and **pt** in the file name and within the transcripts themselves does not represent any restrictions in terms of automatic analyses of the corpus. Converting the **en** and **pt** codes to **eng** and **por** was therefore considered unnecessary and the original codes were maintained in both the file names and transcripts.

purposes. Within the CLAN programme, the selection of the desired files occurs via the 'FILE IN' button which, on being pressed, opens another window displaying all the files available from the specified corpus. In this window it is then possible to transfer the pertinent files to the 'Files for Analysis' box. More details about the procedure for file selection can be found in section 3.1 of the CLAN Manual (MacWhinney, 2014c) and as such will not be repeated here.

As is evident from the above discussion, even the apparently simple process of naming files can have important practical implications. Ideally such decisions need to be made at the outset and will ultimately depend on the particular nature of the corpus and the researcher's objectives in building the corpus. The same applies to the type of coding used throughout the corpus, as will be seen in the following discussion.

3.2 Transcribing and coding the LOBILL Corpus

As mentioned before, the CHAT system offers a standardized format for transcribing all types of conversational interaction, providing options for basic discourse transcriptions as well as more complex phonological and morphological analysis. Apart from the obligatory conventions which need to be followed in order for the CLAN programmes to work, it is up to each individual to choose which further codes need to be inserted into the transcriptions of any particular data set: this will depend on the researcher's aims. As will be seen in this section, several of the codes chosen for the LOBILL Corpus serve very specific purposes and were inserted to enable the investigation of bilingual phenomena.

In a previous version of this chapter a detailed description of how I transcribed and coded the data was so comprehensive that it extended for almost forty pages. Although such detail is necessary for those unfamilar with CHAT conventions and those wishing to build similar corpora, due to textual confines, the decision was taken to remove most of the original description from the body of the dissertation and place it in an appendix (see Appendix B). By doing this more focus could be given to the results of my investigation while still allowing readers to have access to an important part of my methodology. Nevertheless, certain decisions regarding the types of codes used in the LOBILL Corpus need discussing in full here as it is only with the insertion of these specific codes that such a novel investigation of code-switching was made possible. Before turning to this discussion it is important to mention that throughout

this section, all examples are taken from the LOBILL Corpus and are presented in Times New Roman, without quotation marks. This font was chosen for its similarity to the font used in CHAT (ASCII) and quotation marks will not be used as they exert a specific function for analysis purposes.

3.2.1 Coding bilingual data

In Chapter 2, in section 2.2.4.1.2, the coding of bilingual data was discussed with reference to the CHILDES database and the LIDES proposals. It was pointed out that although the LIDES system offered a more consistent way of coding bilingual corpora which would allow for comparative analyses across corpora, there were still simple improvements that could be made to the system proposed. In this section I will detail these improvements and propose additional codes, outlining how they would make the transcription process more effective and increase the potential for automatic analyses of code-switching in electronic corpora.

3.2.1.1 Coding the languages

Of the methods suggested for the notation of bilingual data in the CHAT manual, the one which appears to be most relevant for the current study involves the marking of every single word with the symbol **@** followed by a letter(or letters) which represents the language of the word. For example, the following excerpt from the LOBILL Corpus would be transcribed as shown below (where **@en** represents 'English' and **@pt** represents Portuguese)[15]:

(2a)
*JAM: no@en lá@pt has@en got@en a@en piscina@pt, animais@pt.
*MOT: yeah@en?
*MOT: is@en that@en the@en one@en Meggie@en went@en to@en?
F040: L24

This was the method chosen by the LIPPS group to code their bilingual data (Gardner-Chloros et al, 2000). However, as mentioned previously, two disadvantages of using this type of coding are evident: firstly, the insertion of codes after each word in a transcript is extremely time-consuming; secondly, the transcript becomes more difficult to read. Nevertheless, this method would ensure the complete retrieval of code-switched material for analysis purposes.

[15] Glosses have not been provided here or elsewhere in this section as the focus is on the codes and not the meaning of the utterances.

Through experimental insertion of certain symbols and subsequent testing via the FREQ and KWAL commands (see section 3.4 for details of these CLAN commands), it was discovered that it was possible to economize on the use of symbols without compromising the results outputted by CLAN. If we take the same excerpt, it would now look like this:

(2b)
*JAM: no[@en], lá[@pt] <has got a> [@en] <piscina, animais>[@pt].
*MOT: yeah[@en]?
*MOT: <is that the one Meggie went to> [@en]?

Instead of labelling every single word with the **@** symbol and the language code (**en** or **pt** in this case), angled brackets can be used to delimit sequences of words in the same language. As long as the brackets are immediately followed by parentheses containing the language code, CLAN is able to retrieve the desired material in the same way as if each word had been coded separately. On the basis of the excerpt shown above, one may try to argue that the economy in transcription achieved is minimal. However, if one considers bilingual discourse can often consist of utterances which alternate between languages on an inter-sentential level such as the last utterance in the excerpt), it is then possible to fully appreciate the extent of the economy provided by this improvement. Furthermore, in terms of legibility, the second excerpt also offers a clearer read. In section 3.4 we will see how this type of language coding will allow for both the quantitative and qualitative investigation of code-switching.

It is important to point out that the language coding of forms which exist in both English and Portuguese did not present a problem. For example, in (2b) above, although the written form 'no' also exists in Portuguese (not to express a negative but as the assimilated form of 'em' ('in') and 'o', the masculine definite article), the pronunciation of both forms is quite different: whereas the English negative (as used by JAM in the excerpt) is pronounced [n], the Portuguese preposition is pronounced [nu]. Of course, in the excerpt above, prosodic and semantic clues are also available to help determine the coding of the bilingual 'homograph' 'no' as belonging to English in this instance, and not to Portuguese. Throughout most of the transcription process, the phonological differences between identical English and Portuguese written forms (such as 'zero', pronounced ['zɪərəu] in English as opposed to ['zɛru] in Portuguese

and 'animal', ['ænɪməl] as opposed to [ani'maw]) made the language coding of such forms straightforward. In rare instances, where there was phonological similarity, other clues were used to determine which language code should be applied. For example, in the utterance 'Me dá o homem' ('Give me the man'), based on pronunciation alone, one could theoretically code the indirect pronoun 'me' as either Portuguese or English. However, its position in the utterance and the fact that the remainder of the utterance is in Portuguese makes it very unlikely that the speaker is using the English indirect pronoun here. Therefore the whole utterance would be coded as Portuguese: <me dá um homem>[@pt] (see F107:L17). Such coding decisions are important as they ultimately affect the output provided by the CLAN analyses carried out in this research.

In order to cater for certain forms which could not be coded as belonging exclusively to either English or Portuguese, a special code was devised. As can be seen in the example below, such mixed forms are coded by the @ symbol and then the letters **mf** (mixed form). The underline does not form part of the transcription:

(3)
*JAM: <mas só que só fica um>[@pt] <train+track>[@en] <e um>[@pt]
bonde_track@mf também[@pt] . F062: L199

In this example JAM exploits the English compound 'train+track' by replacing the first element 'train' with 'bonde' (the Portuguese word for 'tram') to become the mixed form 'bonde_track'[16]. By coding all instances of mixed forms (relatively scarce in the corpus) instant retrieval and subsequent analysis of these occurrances is possible.

3.2.1.2 Coding code-switched utterances

In addition to coding the languages, there is another type of coding I would like to propose which would allow for further investigation into the nature of the code-switches. The creation of this specific coding arose as a result of some useful feedback I received at a talk I gave at Lancaster University about the LOBILL Corpus. The question arose as to whether it would be possible to code the direction of the switches i.e. whether the speaker switched from English into Portuguese or vice versa. Seeing the value of being able to investigate this electronically, the CHAT

[16] In the case of established compounds the CHAT convention is to transcribe them by inserting an addition symbol between the elements of the compound (as in 'train+track'). As JAM's use of 'bonde_track' is novel, an underline can be inserted instead of the addition symbol, thereby indicating a non-established compound.

manual was examined to see whether there were any options for achieving such coding. Although no specifically-designed codes were found, it was realised that a special 'postcode' could be created and inserted for each code-switched utterance.

As explained in the CHAT manual (2014b:75), 'postcodes' are symbols which occur in square brackets at the end of utterances and apply to the whole utterance. In the format **[+ text],** a researcher can create their own postcodes according to their specific research aims. In the case of the LOBILL Corpus, the following format was designed in order to code the directional nature of the code-switched utterance: **[+ ep]** for utterances in which the speaker switches from English (**e**) to Portuguese (**p**); **[+ pe]** for when the speaker switches from Portuguese to English; **[+ epe]** for switches which start in English, switch to Portuguese and then switch back to English and so forth. The following examples illustrate the use of these postcodes (note that the underline is not part of the code and serves merely to highlight the postcode):

(4)
*JAM: <just smash them>[@en] <né assim>[@pt] ? [+ ep] F096: L626

(5)
*MEG: <mas <a agua>[//]>[@pt] <the water is very very cold>[@en] ? [+ pe]
F096: L630

(6)
*JAM: <which[//] yes we did some more at>[@en] <eles[//] as crianças todinho do mundo até da Inglaterra>[@pt] <we are ghosts>[@en]. [+ epe] F092: L230

(7)
*JAM: <quan(do)[//] mas quando era um>[@pt] ghost[@en] <era[/] meu nome era>[@pt] <Mister ghost>[@en]. [+ pepe] F092: L286

Occurring after the utterance terminator, the postcode is not followed by any punctuation. It can contain any number of **p**s and **e**s and therefore can cover any number of switches which may occur in one utterance.

It was decided to exclude most proper names from counting as a switch, despite their linguistic coding as English or Portuguese: note the postcodes of the following two examples:

(8)
*MEG: Hamtaro[@pt] <which one wants>[@en] <criançinhas>[@pt]. [+ ep]
F096: L49

(9)

*JAM: <Daddy@p put the>[@en] Palio[@pt] on[*][@en] Vovô@pn[@pt] 's[@en] garagem[@pt]. [+ epep] F015: 140

In the first example 'Hamtaro' is the name of a Brazilian fictional character and, with no English equivalent, cannot count as a switch, In the second example 'Palio' is the model of a car and is not included in the postcode as constituting a switch for the same reason.

However, certain kinship forms and some proper names were counted as codeswitched elements when it was evident that the speaker was making a choice between two alternatives. For example, in the following utterance, JAM says 'Mum'[17] in English before switching to Portuguese and then back to English with the word 'crazy'.

(10)
*JAM: Mum@m[@en] <na Inglaterra eu era muito>[@pt] crazy[@en]. [+ epe] F096: L654

Of course, if JAM only ever used 'Mum' when talking to his mother (and not the Portuguese equivalent 'Mãe', it could be argued that no choice is being made and that, therefore, it would not constitute a switch and should not be coded thus. However, based on simple frequency analyses of these two forms of kinship in the corpus[18], it was possible to determine that JAM did use both forms when addressing his mother and when referring to her in the third person. This was also true of another bilingual kinship form, 'Pai' ('Dad')[19]. However, when addressing or referring to other relatives, a frequency analysis revealed that both children did not vary in the usage of kinship terms: they used 'Avó' and 'Avô' consistently for their Brazilian Grandfather and Grandmother and 'Grandma' and 'Grandad' for their British grandparents. This also applied to uncles and aunts where the kinship form became part of the proper name of the relative, such as 'Auntie_Becky' and 'Tio_Pedro' ('Uncle Pedro').

With regard to proper names, again where a frequency analysis of the corpus showed that two different linguistic forms had been used by the speakers, these forms were coded as a switch in bilingual utterances. Such bilingual options would include for example, whether a speaker chooses to say 'London' or 'Londres', or 'Cathy' or 'Catarina'.

[17] In the LOBILL Corpus, all direct and indirect references to the mother are coded with **@m** as in 'Mum@m' or 'Mãe@m'.
[18] See section 3.3.3 for details of how to carry out such analyses.
[19] All kinship terms referring to the father are coded with **@p**, as in 'Pai@p' or 'Dad@p'

Despite these considerations, most bilingual utterances were straightforward in terms of coding with the postcode presented above. In the case of utterances containing mixed forms (coded with **@mf**), the letters **mf** were simply included in the postcode in the appropriate place:

(11)
*JAM: vou[@pt] press[@en] vou[@pt] pressar@mf <daqui[/] daqui>[@pt] +... [+ pepmfp] F004: L199

In this utterance, JAM appears to create a Portuguese version of the verb **press** by adding the suffix **ar** (the most common infinitive verb ending in Portuguese). The presence of this mixed form is indicated in the postcode.

As will be seen in section 3.3, by coding code-switched utterances with this specifically designed postcode, their retrieval in the corpus is facilitated and comparisons across speakers can be easily made.

3.2.2 Coding tag questions

When building a corpus, not all decisions about coding will necessarily be made a priori: it is often the case that through the actual transcription process itself, the need for other codes to be inserted arises. In the case of the LOBILL Corpus, it was perceived that the use of English tag questions by JAM in monolingual utterances appeared to diverge from what would be considered the norm. Frequently, he would use 'is it?' or 'isn't it?' where the use of a different auxiliary would be the expected form, as in the following example:

(12)
*JAM: <it stops the blood going out, isn't it>[@en]? F076: L305

Although in some varieties of English it might be the norm to use such an all-encompassing tag question, given the fact that such usage was not noted while transcribing the other informants' utterances (those of both bilinguals and monolingual English speakers) it does not seem likely that JAM was following a linguistic norm. However, if we consider that the Portuguese equivalent of 'isn't it' ('né') functions as a generic tag question, one is led to speculate that JAM might be superimposing the linguistic norm of Portuguese on his English tag question usage.

To be able to investigate the occurrence of this phenomenon in the corpus, especially in code-switched utterances, the decision was taken to code all instances of tag questions, whether they be in English or Portuguese. The format chosen can be seen in the following two monolingual examples:

(13)
*MEG: <he's as well talking in Portuguese, <isn't he>[@tq]>[@en]? F025: L282

(14)
*PAI: <então ótimo vamos livrar daquele lá, <né>[@tq]>[@pt]? F039: L223

Enclosed in angled brackets and followed by the symbol **[@tq],** CLAN would thus be able to retrieve these structures wherever they occur in the data, including those in code-switched utterances.

3.2.3 Coding extra-linguistic information

All of the codes thus far discussed involve their insertion in the speaker utterances themselves (referred to as the 'main lines' in the CHAT system). To enrich a corpus even further more information can be added beneath each utterance on the 'dependent lines/tiers'. Unlike the first two components of the CHAT system (the headers[20] and main lines), the use of the dependent tier is completely optional. Although there are no requirements for a researcher to provide information on the dependent tier, it is evident that by doing so the corpus becomes a much richer resource for investigation purposes. CHAT provides several options for coding but also allows for the creation of novel codes for specific purposes.

All dependent tiers start with the percent symbol **%** which is then immediately followed by a three-letter code in lower-case and a colon. After a tab the metalinguistic information is then included. It is possible to insert as many dependent tiers as desired for each main line utterance, one below the other. A complete list of the codes can be found in the CHAT manual (2014b: 78-84). While accepting that the more coding there is, the richer the corpus will be for general investigation purposes, the time limitations of a research project mean that the extent of the coding must be determined by the specific research questions. For this reason, it was decided to make use of the following three dependent tier codes in the LOBILL Corpus: **%add:** ('addressee'), **%com:** ('comment') and **%err:** ('error'). Of these three, the addressee

[20] Information about headers can be found in Appendix B.

code proved to be very important for analysis purposes and its use will be discussed in detail below.

3.2.3.1 The dependent tier code **%add:**

As explained in 2.2.4.1.2, a crucial variable to consider when investigating code-switching is the role of the interlocutor: the addressee's linguistic background (such as their degree of bilingualism, their ideological attitudes to language, their familiy and social linguistic practices) can all have significant effects on the code-switching practices of the speaker and vice versa. Without coding the addresse of each utterance in the corpus, these aspects cannot be fully explored and I pointed this out as one of the oversights of the proposals put forward in the LIDES system. Thus a further improvement that is proposed by the present research is the mandatory inclusion of addressee coding in the corpus.

To illustrate this type of coding, and its importance, I will discuss the short excerpt shown below in which JAM is talking to his mother about what birthday present to buy for his friend:

(15)
*JAM: buy[@en] +...
%add: MOT
*MOT: <a what>[@en]?
%add: JAM
*JAM: <two beyblades>[@en][= whispers].
%add: MOT
*MOT: <two beyblades>[@en] +!?
%add: JAM
*MEG: no[@en]!
%add: MOT JAM A
*MOT: why[@en]?
%add: JAM B
*JAM: <because one for me (a)n(d) one for>[@en] Rafa[@pt].
%add: MOT
*MOT: <it's not your birthday>[@en].
%add: JAM
*JAM: <I know but>[@en] <dia das crianças>[@pt]. [+ ep]
%add: MOT C
*JAM: <buy[/] buy me one>[@en] <dia das crianças>[@pt] <and one for>[@en] Rafael[@pt]. [+ epe]
%add MOT D
*MEG: uhuh[= shakes head].
%add: JAM
*MEG: <James@pn, tu não entende dinheiro>[@pt].
%add: JAM E

*MEG: <din(heiro)[/] dinheiro é importante>[@pt].
%add: JAM F F080: L445-471

JAM suggests his mother buys two beyblades (a modern-day spinning top) and when she expresses puzzlement about why she should buy two, MEG, who is present at the table, protests. Her use of 'no!' is clearly addressed to both her brother and her mother and thus both addressees appear coded on the dependent tier (see A). The mother's subsequent 'why?' is addressed to JAM (B), although without specific contextual information, her question could easily have been interpreted as being addressed to MEG. Here, the use of the **%add** code clearly helps to clarify the addressee.

In C and D JAM is addressing his mother and we see that he switches to Portuguese to insert 'dia das crianças' ('children's day') in an otherwise English utterance. When MEG again interrupts, she does so in Portuguese, this time addressing JAM, saying 'tu não entende dinheiro' ('you don't understand money') and then 'din(heiro)[/] dinheiro é importante' ('money is important') (see E and F).

From this short interaction, it is possible to see how important the question of addressee is when examining the language use of bilingual speakers. Despite being addressed exclusively in English by his mother, JAM uses a Portuguese phrase while MEG addresses her brother purely in Portuguese.

As mentioned before, the interlocutor variable should be considered a crucial issue for anyone wishing to research bilingual language use and code-switching. In an electronic corpus, such as the LOBILL Corpus, the use of an addressee code provides an effective way of investigating the relationship between each speaker's language use and his/her interlocutors. This will be shown throughout Chapters 4 to 7 in the discussion of the results of both the quantitative and qualitative analyses performed on the corpus.

Although the addressee code was inserted under every utterance in the corpus, the second code, **%com:** , was used on a more ad hoc basis, as shown below.

3.2.3.2 The dependent tier code **%com:**

As can be seen in the following examples, this is a multi-purpose code which can provide a variety of comments.

(16)
*MOT: <alright, go and see Daddy@pn ‘s photos>[@en].
%com: father enters room F080: L556

(17)
*MEG: +< <dois carros>[@pt] .
%com: Meggie is reminding James what he got for Christmas F094: L200

(18)
*MOT: <so tell us, what[/] what is the news>[@en]?
*MEG: <well Milly the guinea_pig, she's>[@en] +...
*MOT: what[@en]?
*MEG: <er (..) with babies in her tummy>[@en].
%com: Meggie clearly wants to say pregnant but can't remember the word in English and gets around it by paraphrasing F096: L21-30

For researchers wishing to be more specific in classifying this type of information, the following codes would provide this specificity: **%act:, %exp:, %par: and %sit:** (see the CHAT manual for more details). However, as this is not a necessary goal for the current project, the **%com:** code was used to englobe any pertinent extra-linguistic information.

3.2.3.3 The dependent tier code **%err:**

Detailed instructions on the use of CHAT error coding can be found in section 15 of the manual (pp 100-105). Before I discuss the choices I made regarding error coding in the LOBILL Corpus, it is important to draw attention to the term 'error' itself, as it can be the cause of much debate among researchers in the field of error analysis.

The fact that in the CHAT manual MacWhinney himself does not discuss, or justify, the use of this particular term implies that, for him at least, it is a given that the word 'error' can be used whether one is talking about child or adult language or whether monolingual or bilingual language use is being discussed. Indeed, recent literature in the field does appear to reflect this widespread acceptance of the term 'error'. With regards to studies of monolingual children, all the following make reference to language 'errors': Räsänen et al (2013) on Optional Infinitive (OI) errors; Ambridge (2013) on linguistic generalizations; Dodd (2013), Brosseau-Lapre (2013) and Gildersleeve-Neumann & Goldstein (2014) on speech disorders; and Jaegar (2013) on slips of the tongue. Examples of studies focussing on 'errors' in second-language learners include Coyle & Roca de Larios' research on error correction

strategies in second language acquisition (2013) and Granger's investigation of errors in the FRIDA (French Interlanguage Database) corpus. As for studies of errors in bilinguals, in Gollan et al's study of adult bilinguals (2014), in Gagarina's study of a Russian-German bilingual child (2013) and in Gillam et al's study of Specific Language Impairment (SLI) in bilingual children, the term 'error' is used without reservation. Although in the latter study, the authors do discuss the difficulties in distinguishing 'developmental' errors from those caused by SLI, they do not have a problem with the term 'error' itself when describing their research. Indeed, in James' 2013 book on error analysis, he comments that by simply using the word 'error', we are in no way jumping to any conclusions: “The explanation (or 'diagnosis') of a unit of learner language that does not match its equivalent in the TL [Target Language] is in no way prejudged by the simple act of calling it an error.”(2013:17).

It appears that the issue with 'error' is methodological rather than terminological. How can errors be identified? What is the best procedure to follow? Is it appropriate to compare a child's developing language to an adult? Should we use monolingual norms as a basis for studying a bilingual's language errors? In his discussion of the problems of error identification in L2 learners, Lennon points out that even using a monolingual norm is problematic, as 'considerable variation is to be found even among native speakers' (1991:181). For Corder, error identification and interpretation (of L2 leaner errors) is made easier if speakers are available for consultation - by being able to consult with the informants themselves, a researcher is able to make 'authoritative' interpretations (rather than just 'plausible' interpretations) of their communicative intentions (Corder, 1981). In my study, my familiarity with the bilingual informants and my immersion in the bilingual language context means that I was in a very good position to make 'authoritative' interpretations of the speakers' intended meanings. Although in practice it was seldom necessary to consult with any of the speakers of the LOBILL Corpus in order to identify an error appropriately, there were occasions where clarification occurred within the recorded dialogue itself, as can be seen in example (19) below.

At this point it is appropriate to leave aside any further theoretical discussion of issues related to the field of error analysis and return to the specifics of my study, in which error analysis plays a small, albeit significant, role. Using examples from the LOBILL Corpus, I will now illustrate how error coding can be carried out with the

CHAT system and discuss the choices I made with regards to the level of error coding in my corpus.

The error code on the dependent line (**%err:**) is used in conjunction with the **[*]** symbol which is inserted on the main line immediately after the error. The following example shows what type of entry the **%err:** code normally takes:

(19)
*JAM: <I just fell off[*]>[@en].
%add: MOT
%err: off = over;
*MOT: <fell off what>[@en]?
%add: JAM
*JAM: <I fell off[*]>[@en].
%add: MOT
%err: off = over;
*MOT: <you mean you fell over>[@en].
%add: JAM
*JAM: yes[@en].
%add: MOT F078: L79-90

On the dependent line the error is followed by an equals sign and then the target form. Where the error involves more than one word, angled brackets should be used to show the extent of the error, as the following example demonstrates:

(20)
*JAM: <and look Mum, I think (be)cause downstairs <it has>[*] a fire>[@en].
%add: MOT
%err: it has = there is; F079: L316

If more than one error needs to be coded on the same line, they should be separated by a semi-colon.

The examples above show the simplest form of error coding which does not aim to indicate the classification of the error. CHAT does provide more specific codes which can be used to detail the type of error and these can be found in the manual. Although it is not the aim, and beyond the scope of the present research, to perform an error analysis on all of the errors identified in the LOBILL Corpus, it was predicted that in the investigation of code-switching in the corpus, certain errors may arise as a result of the interaction of a speaker's two languages. Being able to locate and examine such errors meant that some form of error coding needed to be undertaken. However, would it be viable to use the detailed error coding options provided by the

CHAT system? The following discussion will reveal how complex this question proved to be.

In order to show how challenging the classification of errors in bilingual speech can be let us examine the following utterance, where JAM is telling his mother about an incident with his sister:

(21a)
*JAM: <I just telled <buy the beyblade>[“] and she hit me>[@en] !
%add: MOT F080: L509

The material within the angled brackets followed by [“] indicates a quote and JAM is complaining that as a result of saying this his sister had hit him. On first examination it appears that there are two errors in this utterance: JAM is using 'telled' instead of 'told' and he also omits the prounoun 'her'. This being the case, the utterance would be transcribed as follows:

(21b)
*JAM: < I just telled[*m:d] 0her <buy the beyblade>[“] and she hit me>[@en] !
%add: MOT
%err: telled=told

The asterisk is followed by **m** which means that the error is morphological and **=d** which means that JAM has overregularized the past tense ending. To indicate a missing word the numeral **0** is used, followed by the missing word, in this case 'her'. However, on examining the utterance within its wider context we learn that JAM really meant to say 'said' and not 'told', the quoted speech having originally been directed to his mother (i.e he was not telling his sister to buy the beyblade). This means that the underlying error is semantic in nature and not simply morphological. Support for this conclusion can be found in a simple concordance analysis using the key words 'tell' and 'say' (and their derivatives): the output showed that JAM frequently used the verb 'tell' when the verb 'say' would have been more appropriate. If we consider that a single Portuguese verb, 'dizer', is usually used to express both English verbs, it is probable that through a process of linguistic transference, JAM is overgeneralizing the use of the English verb 'tell'. This second analysis has implications for the error coding of this utterance which would now look like this:

(21c)
*JAM: <I just telled[*s:r][*m:d] <buy the beyblade>[“] and she hit me>[@en]!
%add: MOT
%err: telled=said; telled=told

After the first asterisk we now find **s** which means the error is semantic and **r** which means that it is a related word. On the dependent tier we also find 'said' as the target form. Although the error is essentially semantic in nature, it would also be important to code the use of 'telled' instead of 'told' as this would still constitute a formal error worthy of tracking in the longitudinal corpus.

As is evident, error analysis is an extremely complex area, even more so when it involves examining errors found in the speech of bilingual children. The limitations of the current research project in terms of time and scope mean that the LOBILL Corpus does not contain the detailed error coding illustrated above. Rather than classifying errors, it was decided to merely indicate the location of an error on the main line with the asterisk code [*] and provide the error and target form, if known, on the dependent line following the code **%err.** It was latterly discovered that CHAT also allows for the insertion of comments on this dependent line and therefore whenever possible I added explanatory notes as to the possible origin of the error. Thus I was able to provide more information about an error without having to resort to the more complex error coding detailed in the CHAT manual.

Opting for a more simplistic method of error coding meant that minimal additional time would be needed to include the codes during the transcription process. However, it would still mean that a detailed error analysis could be performed at a later date: every error in the corpus would now be instantly retrievable and more detailed codes could be inserted efficiently (ie there would be no need to trawl through pages of transcript).

Such foresight in coding the LOBILL Corpus means that its utility as a rich data base will go beyond the current research. Once contributed to CHILDES, the corpus will represent an useful resource for those researchers working in the field of error analysis, partly due to the ease with which errors can be located and subsequently analysed. The additional comments inserted on the dependent line also have the potential to provide other researchers (who may not speak Portuguese) with certain insights into the data which might aid their eventual interpretation of the errors.

This section (3.2) provided details of the specific codes used in the transcription of the bilingual data in the LOBILL Corpus and discussed the rationale behind these choices. The importance of these codes for analysis purposes will become apparent in the next section which will specify the CLAN commands used to investigate code-switching in the data.

3.3 Analysing the LOBILL Corpus

So far in this chapter on the methodology of the present research, the nature of the LOBILL Corpus has been described and the process of data collection, transcription and coding has been detailed. This section will now examine how the corpus was analysed using the CLAN (Comptuerized Language ANalysis) programme. Again, due to textual constraints, the original version of this section was severely edited and, as such, more general information about the use of the CLAN tools will not be discussed here. Interested readers may consult the CLAN manual (MacWhinney, 2014c) for further details and for the full gamut of commands available for linguistic investigation.

It is important to mention here that although this section also originally set out a plan of analysis of the LOBILL Corpus, as the investigation progressed it became apparent that it was becoming more and more data-led. That is, the results from one analysis would reveal something which merited further investigation, therefore leading to further analyses not predicted in the original plan. This unpredictable aspect of the methodology has meant that it is not logical to present a definitive plan of analysis in this section – decisions about subsequent analyses were dependent on findings and the discussion of these findings falls within the remit of the following chapters. Therefore, in this section my focus will be on introducing the specific commands used throughout the investigation and explaining how their potential for analysis was maximised through the coding described in the previous section. I will describe them in alphabetical order and, where appropriate, provide examples of command lines. Before looking at each command on an individual basis, however, I will first make some general observations regarding the construction of the command lines.

3.3.1 Constructing command lines

CLAN command lines are constructed by combining the different commands with different types of search strings: they may be strings of letters (i.e. words) or strings made up of other characters such as the symbols used to transcribe in CHAT format or especially designed codes used to annotate the corpus. In most cases the search for a particular string of characters is achieved by enclosing the target word, symbol(s) or code in double quote marks and preceeding it with +**s** (s standing for 'string'). For example, to search for occurrences of the word 'going', the retracing symbol ([//]) or tag questions (coded with [@tq]), the search strings used would be as follows: +s"going", +s"[//]" and +s"[@tq]".

Many of the searches carried out in this study involved the use of the asterisk symbol (*), which in CLAN can function as a 'wild card' character. Used to represent any number of characters, a string such as +s"chang*" would enable the programme to search for all forms of change such as changes, changed or changing. As will be seen in the discussion of the results, this metacharacter was used to search for code-switched material: its use in the string +s"[+ *]" means that all combinations of letters following the + in the CS postcodes (such as [+ ep], [+ pep] and [+ epepepe]) are included in the search.

In 3.2.3.3 we saw that the asterisk symbol is also used to code errors on the main line. Despite its use as a wild card metacharacter as described above, the search for the error codes is still straightforward: the search string +s"[*]" will find all of the error codes in the specified input. If the square brackets were removed, however, (as in +s"*"), the output would effectively then include all the material for that specified speaker.

As seen above, the use of brackets in a search string can have an important effect on what a particular command is instructed to search for. Some searches are also affected by the types of brackets used. For example, when using FREQ (see 3.3.3 below) the output for the search string +s"[@tq]" (using square brackets) consists of the total number of tag question codes found in the data while the output for the search string +s"<@tq>" (using angled brackets) is a frequency word list of all the tokens coded with [@tq]. For my study both types of output were valuable.

Of course most searches are performed on a specified speaker's utterances (as opposed to on the corpus as a whole) and thus most command lines will need to include this information. Rather than use the +s switch, in order to select a speaker's main lines the +t switch is used (t standing for 'tier'). Thus +t*JAM will tell the

programme to examine only those tiers beginning with the speaker code *JAM , excluding by default all other speaker tiers. Due to the addressee coding inserted in my corpus, I am able to specify even further and instruct the programme to only look at those utterances addressed to a particular interlocutor. To do this, the +t switch is used to instruct the programme to also look at addressee tiers (as in +t%add), and the +s switch is then used to specify which particular addreesse I am interested in (for example, +s"MOT"). The facility to be able to take the variable of addressee into account was crucial for the analysis of the code-switching practices of the informants of this study.

Before moving on to the specifics of the CLAN commands themselves, it is useful to mention two other switches which are often used in the command lines and which perform the same function across the commands. One of these is +u, which instructs the programme to merge the results of the analyses carried out on each file - the default is to provide separate output for each file. The other is +f, which sends (and saves) the output to a separate file, instead of it being shown in the output window. Useful for the filing and tracking of results, the use of this switch becomes obligatory in cases where the output is too extensive to be shown in its entirety in the CLAN window. By using the +f switch, the saved results do not suffer this truncation.

Although there are several other switches, their function may vary according to the command used or even be unique to a particular command. Therefore, they will be specified in each of the following separate sections which detail the functioning of the five commands used in this study.

One last observation about the construction of command lines is that each one must begin with the name of the command, as shown in the examples below. After typing in the command, however, all the other elements which make up the command line (the specific strings and switches) can be placed in any order.

3.3.2 COMBO

As the name suggests, COMBO is designed to search for combinations of words and outputs them in the form of utterances. For example, in order to search for utterances containing the cluster “going to” the following command line would be typed into the

CLAN commands window[21]. Note that the @ symbol is not typed but appears automatically after the files have been selected from the drop-down File menu:

(C1) **combo @ +s"going^to" +u –w2 +w2**

The words of the cluster need to be linked by the symbol ^ and the **–w** / **+w** switches can be used to expand the number of utterances before (**-w**) and after (**+w**) the key cluster. Below is an example of what part of the output looks like: the target clusters are automatically numbered by COMBO[22]:

*GRA: <present time>[@en] .
*GRA: <it's like Christmas time>[@en] .
*MEG: <Mummy@pn # you're gonna[: (1)going (1)to] get his xx and he's
gonna[: (2)going (2)to] be all black>[@en] .
*MOT: <xx I'll put>[@en] +...
*MEG: Mummy@pn[@en] .

File names and numbers (not shown here) are provided along with the output and this is clearly very useful as due to the longitudinal nature of the data in the LOBILL Corpus, it is important to be able to factor in the variable of age/time of occurence. Also included at the bottom of the output is the total number of times the key string is matched in the data selected for analysis. The inclusion of this total means that it is unnecessary to count up the number of occurences in the output manually.

3.3.3 FREQ

In its simplest form this command produces a list of all the words in a specified file, or group of files, indicating their frequency and calculates a type-token ratio. This calculation is achieved by taking the total number of different words used by the speaker (types) and dividing it by the total number of words used (tokens). This result gives an indication of the lexical diversity of a speaker (but see discussion in 3.3.5). While the output from a frequency analysis can be valuable in purely quantitative

[21] Note that in this and the following sections, the letter and number (C1) do not form part of the command line and are used for reference purposes only. The initial capital letter stands for the particular command being used, in this case COMBO.

[22] The numbering is not accumulative across the different utterances and therefore most target clusters will be numbered (1) as they occur only once in an utterance. In the example provided there are two occurrences of the key cluster in a single utterance and thus they are numbered (1) and (2).

terms (such as the total numbers of words), the actual word lists themselves provide a rich source of data for analysis.

By default FREQ ignores headers, dependent lines and the symbols **xxx** and **www** (used to transcribe uninteligible speech and non-transcribed material respectively). It also excludes any words which begin with the following codes: **0, &,** **+, -** and #. Examples include **0it** ('it' being marked as omitted) and **&blee** (a phonological fragment of a word). When it comes to the treatment of assimilations (such as gonna and wanna) or shortenings (such as can't and don't), if these forms are followed by their full forms in square brackets (e.g. gonna[: going to] and don't[: do not]), FREQ will automatically perform its analyses on this material instead of the form preceeding it. There were two reasons why I considered this 'text replacement' selection to be undesirable for my analyses. Firstly, I wished to analyse the material actually produced by the speakers and considered that forms such as 'gonna' and 'going to' should be treated as separate forms. Secondly, any inconsistencies in the transcription of the full forms (for example, not using this type of notation for every single occurrence of gonna) would ultimately effect the results. I chose, therefore, to make use of the +**r5** switch which overides the default, thereby forcing FREQ to only select the original forms for its analyses (i.e. ignoring any 'replacement' material). It is pertinent to mention here that to enable more reliable triangulation of the results of the frequency analyses with those of the other CLAN commands, this switch was also used for all of the VOCD and WDLEN analyses (see sections 3.3.5 and 3.3.6). In a similar fashion, there was a particular search string which I decided to include in all the command lines with FREQ, VOCD and WDLEN. This string is discussed below

While the language coding of the files was mostly straightforward (in terms of deciding which words belonged to each language), I was faced with an important decision regarding the treatment of what I termed 'non-words' such as 'err', 'erm', 'urgh', 'mmm' and 'ssshh', of which there were a total of 63 different types![23] If I simply ignored them and left them without any language coding at all I predicted that I might find it difficult to exclude them at a later date if desired.

In order to exclude further groups of words from a frequency analysis, CLAN allows the researcher to simply list these words in a specially created file which, when then used with the switch -s, are excluded from the input. Choosing this option,

[23] See Appendix C for the complete list of non-words.

I therefore decided to code these non-words according to the language immediately surrounding the non-word but then include them in a special file I named **nonwords**, with the extension **.cut**. I am then able to instruct the CLAN programmes to remove these words from analyses by simply using the string **-s"@nonwords.cut"** in the command line. I considered this facility to be important as the frequency of these non-words in the data could have the effect of watering down the contrasts that I expected to find in the data regarding the contribution of both languages to code-switched utterances. By removing such distractors, patterns would become easier to identify, resulting in more effective comparisons across the output of different speakers. Again, as with the +**r5** switch, for reasons of consistency this particular string was used with VOCD and WDLEN, as well as with FREQ.

The following four example command lines show how the different elements mentioned above (the strings and switches) can be incorporated into a frequency analysis in the LOBILL Corpus. In the first example FREQ is instructed to examine only those utterances pertaining to JAM (+**t*JAM**), to remove all non-words (-**s"@nonwords.cut"**), to select the original forms of assimilations and shortenings as opposed to the full forms found in square brackets (+**r5**), to merge the results (+**u**) and put them in order of frequency (+**o**).

(F1) **freq @ +t*JAM +u +o -s"@nonwords.cut" +r5**

By adding the string +**s"[+ *]"** (see second example below), the command will then perform its analysis on only those utterances followed by a postcode, which in the LOBILL Corpus relate to those containing code-switched material.

(F2) **freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" +r5**

In order to then output separate word lists for the Portuguese or English material contained in JAM's CS utterances, one would then include further search strings, as can be seen in the following two command lines:

(F3) **freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" -s"<@en>" +r5**
(F4) **freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" -s"<@pt>" +r5**

By using the string **-s"<@en>"** , FREQ is being instructed to remove all the English material from the analysis (leaving the Portuguese material) and by replacing **@en** with **@pt**, it is the Portuguese material which is excluded (thereby leaving the English tokens). In the following paragraph I explain why I chose to use this 'exclusion' method (with **-s**) rather than directly selecting the language I wanted to analyse (with **+s**).

When performing initial exploratory frequency analyses with the strings **+s"<@en>"** and **+s"<@pt>"** I noticed the presence of undesirable items in the resulting frequency word lists. These items consisted mainly of codes (such as [/], [//], [//], [@tq], ["], and [*]) but also included text replacement items (for example, [: going to] and [: want to]). The presence of these items in the lists was clearly undesirable as they did not represent words spoken by the informants and their inclusion would therefore skew the word frequency results. The addition of the +r5 switch in the command lines did not effect the removal of the full forms: it may be that in this case the replacement switch was overridden by the use of the initial **+s** switch with the language codes. In subsequent experimental analyses I found that by using the language code with the **-s** switch, the codes were no longer present in the lists. And this time when I included the +r5 switch in the command lines, an examination of the word lists revealed that it had performed its function by counting the original forms (and not the replacement forms found in brackets). Although it was necessary to use the exclusion method with the language codes, the use of the string **+s"[+ *]"** (to select all utterances coded with the CS postcode) offered no such problem: experimental analyses showed that FREQ automatically ignored the symbols, thereby excluding them from the output. And whereas the switch **+r5** proved not to function with the **+s"<@en>"** or **+s"<@pt>",** with the **+s"[+ *]"** switch it maintained its normal functionality (of forcing FREQ to select original forms and not replacement material). In corpora containing little or no additional coding, it is likely that these methodological considerations related to the functioning of the FREQ command would not be an issue. However, the discussion above illustrates how important it is to verify, through experimental analyses, the effect different switches and search strings may have on the frequency output.

The effect that the use of different brackets in a search string (angled as opposed to square) may have on frequency output has already been mentioned (in 3.3.1). For details on the specific search strings where I exploit this difference, the

reader can refer to the Table 5 shown in section 3.3.7 which summarises all of the strings used for the analysis of the LOBILL Corpus.

As will be seen throughout the discussion of the results, FREQ proved to be extremely useful in the analysis of code-switching in the LOBILL Corpus; indeed, it was often via initial analyses with FREQ that interesting linguistic phenomena worthy of further investigation were first revealed. However, it is again important to highlight that it is only due to the specific coding of the LOBILL Corpus (as described in section 3.2) that the results of the frequency analyses were able to reveal so much about each bilingual speaker's language use.

In order to specify the input for FREQ, it was often necessary to use another command, called KWAL, which will now be discussed in the next section.

3.3.4 KWAL

KWAL means 'Key Word And Line' and is the CLAN command used to search for specified strings in the data which are then outputted in the form of concordances (matching lines, or utterances, containing the key word). This is the command which is able to reveal more about the patterns noted in analyses with FREQ. However, another important function of KWAL is to prepare data for a subsequent analysis with one of the other CLAN commands. For example, in the command line shown below KWAL is used to select certain utterances and put it in the correct format so FREQ can then perform analyses only on that selected data.

(K1) **kwal @ +t%add +t*JAM +s"PAI" +d +u | freq +s"[+ *]" -s"<@pt>" -s"@nonwords.cut" +o +r5**

There are two major parts to the above command line. The first part instructs KWAL to select all of JAM's utterances (**+t*JAM**) that are addressed (**+t%add**) to his father (**+s"PAI"**), merge the results from each file into one output (**+u**) and strip them of any extra information (such as line numbers etc)(**+d**) so that the data is then in the correct format to be analysed by the second command. In this case the second command is FREQ but other commands could be used on this prepared data (see sections 3.3.5 and 3.3.6).

The upright line (|) separates the two parts of the command line and tells CLAN that a further analysis is going to be performed. In this case, FREQ will select

only code-switched utterances (**+s"[+ *]"**) and then remove all of the Portuguese words from the list (**-s"<@pt>"**) and any non-words (**-s"@nonwords.cut"**), thus leaving only those English words occurring in the mixed utterances, in order of frequency (**+o**). The resulting output would therefore contain a frequency word list of only the English words which JAM uses when code-switching with his father.

Apart from using KWAL to specify the input for a subsequent analysis carried out by another command, it was also used to search for particular strings in the data. As mentioned before, these strings could be combinations of letters (i.e. words) or particular transcription symbols or codes (such as **[/]**, which codes repetitions). For example, the following command line would output all concordances where JAM used the conjunction 'but' in CS utterances addressed to his father:

(K2) **kwal @ +t%add +t*JAM +s"PAI" +d +u | kwal +s"[+ *]" +s"but" +d**

Again, focussing on JAM's utterances addressed to PAI, in the command line below KWAL would search for all code-switched utterances (**+s"[+ *]"**) which contain reformulations (**+s"[//*]"**):

(K3) **kwal @ +t%add +t*JAM +s"PAI" +d +u | kwal +s"[//*]" +s"[+ *]" +d**

This analysis would allow for the investigation of a relationship between code-switching and reformulations. By changing the speaker and addressee variables in the first part of the command line, the same analysis could be carried out on other speakers in the corpus and the effect of addressee on the extent and nature of reformulations occurring could be investigated.

Due to the insertion in the corpus of codes to label tag questions (see section 3.2.2), this is another linguistic phenomenon that can be investigated. By typing in the following command line, KWAL would search for all code-switched utterances (**+s"[+ *]"**) which contain tag questions (**+s"[@tq]"):**

(K4) **kwal @ +t%add +t*JAM +s"PAI" +d +u | kwal +s"[@tq]" +s"[+ *]" +d**

By removing the instruction to KWAL to select only code-switched utterances (i.e. taking out **+s"[+ *]"** from the command line) all instances of tag questions, whether in

monolingual English or Portuguese utterances or in code-switched utterances would be found. Such an overiew of tag question usage by each speaker would aid interpretations of the more specific results.

A similar approach can be used when investigating the occurrence of errors (coded by **[*]**) and reported speech/metalinguistic comments (coded by **[“]**). KWAL can be used to find the above in code-switched utterances which can then be compared to their occurrence in both monolingual English and Portuguese utterances. As can be seen in the following command lines the search for errors is achieved by adding the search string **+s"[*]"(**the first command line**)**, and for reported speech/metalinguistic comments the search string is **+s'[“]'** (the second line)[24].

(K5) **kwal @ +t%add +t*JAM +s"PAI" +d +u | kwal +s"[*]" +s"[+ *]" +d**

(K6) **kwal @ +t%add +t*JAM +s"PAI" +d +u | kwal +s'[“]' +s"[+ *]" +d**

In the first case, by analysing the concordances outputted by KWAL it would be possible to investigate the relationship between errors and code-switching in the speech of the siblings. Questions such as the following could be asked of the data: Do errors occur more frequently when the informants are code-switching?; What is the nature of these errors?; Are there significant differences between the siblings' errors?; Are errors acompanied by reformulations? If so, do they occur in the same language or is there a switch to another language?

In the second case, by examining all the concordances which contain reported speech or metalinguistic comments, one could ascertain how often these phenomena are realised by a switch to another language. Comparisons could be made across speakers and parental influence could also be factored into the analysis. Again, by removing from the two command lines the string which is used to specify the selection of code-switched utterances, KWAL would output all instances of errors and metalinguistic references found in the data for a particular speaker, thus providing an overall picture of their occurrences and consequently allowing for comparison of patterns across mono and bilingual speech.

It will be the case for some analyses that in order to interpret the output more effectively the examination of a list of single concordances will not suffice. For

[24] Note the use of single quote marks instead of double quote marks for this search string.

example, evidence for the correct interpretation of the use of a particular word or phrase in a different language by the siblings may only be found in a wider linguistic context. It may be that JAM, or MEG, inserts a particular English word in an otherwise Portuguese monolingual utterance following their father's lead in a previous utterance. Thus it is important to be able to ask KWAL to provide a wider linguistic context where necessary. As shown earlier, in 3.3.2, this is done very easily by using the switch **-w** / **+w** and inserting the number of utterances required. An alternative method is to refer back to the original file, accessed quite easily via the CLAN window.

In the discussion of the results, the importance of KWAL to this study will become evident. However, it is my use of two other commands, VOCD and WDLEN, which perhaps provides the most original contribution of this study in terms of methodology. These will be presented in the following two sections.

3.3.5 VOCD

The VOCD command is used to measure vocabulary diversity. Before showing how I used this command to investigate code-switching, certain issues relating to the measuring of lexical diversity need to be touched upon.

First of all, a case needs to be made for the use of VOCD as oppposed to the traditional Type/Token Ratio (TTR) measurement which is already automatically calculated by FREQ when used to output word frequency lists for a given speaker. Although the TTR value has in the past been used to give an indication of a speaker's lexical diversity (in a particular transcript or group of transcripts), the disadvantage of using such a measure is that comparisons across samples containing different numbers of tokens cannot be made. This is because as sample size (i.e. the number of tokens) increases, TTR scores will invariably become lower: a speaker will get to the point where he is not using any new types, and further tokens in the sample are repetitions of his active vocabulary. This thus results in a decline in TTR scores as the sample size increases. The programme VOCD was developed to overcome this problem and offers a reliable measure of vocabulary diversity (D) which can be used to make comparisons across different data sizes. Details on the validation of the mathematical model which gave rise to VOCD can be found in Malvern et al's monograph (2004) and for other studies which have made use of and appraised the D measure, the reader can refer to Durán et al (2004),

McCarthy & Jarvis (2007), Richards & Malvern (2007), Treffers-Daller (2009a, 2010) and Toruella & Capsada (2013).

In order to illustrate the types of values provided by VOCD, it is useful to briefly mention Durán et al's study (2004) in which they calculated D scores for different cohorts of subjects. In their spoken data, the mean D scores for children ranged between 14.8 at 18 months to 64.02 at 5 years old. The mean D scores for their teenage learners of French and adult second language learners of English were very similar: 56.28 and 56.58 respectively. With regards to written data, they found that the range of D scores for academic writing was between 69.74 and 119.20. Although the D scores resulting from my study will not be compared to those shown above, they provide an indication of how to interpret D values at a basic level - that the higher the D score the higher the lexical diversity.

An important issue raised by Durán and colleagues was how to treat inflected forms when measuring lexical diversity. In their child data, they chose to strip off regular inflections and base their count on stem forms, so, for example, 'fall', 'falls' and 'fell' would count as two types. They reasoned that this was the best method to avoid confounding lexical diversity and the development of morphology (ibid, p.228). Treffers-Daller (2013) and Treffers-Daller and Korybski (2015) recommend a more rigorous approach when using measures such as *D*. They say that data should be 'carefully cleaned and lemmatized' (ibid, p.32) in order to avoid D scores which have been inflated by morphological inflections. In Treffers-Daller's 2013 study of L2 French learners, for example, a speaker producing all the following 9 forms 'cherche', 'cherchons', 'chercha', 'ils', 'il', 'tous', 'tout', 'la' and 'le', would have the same number of types (4) as an L2 speaker only able to produce 'cherche', 'il', 'tout' and 'le'. I would argue that in this case such a method would eliminate the differences in diversity that might be present in the productive language of different levels of L2 learners – despite producing a greater number of diverse forms, more advanced learners would achieve similar D scores to less advanced learners. If I were to 'clean' the data in the LOBILL Corpus in the same way before using the VOCD programme, I would be eliminating any opportunity to use D scores as a way of capturing differences in lexical diversity between bilingual siblings whose language development in both languages is at different stages.

Is is when we wish to make comparisons across typologically different languages that the method of lemmatization of data becomes more relevant. In the

other study mentioned above (Treffers-Daller & Korybski, 2015), the authors demonstrate that such treatment of the data reduces the effect of inflectional differences on lexical diversity values, thus enabling better comparisons to be made across Polish, a highly inflected language, and English. However, they do admit that it is 'not possible to completely eliminate all differences between the two languages in this process' (ibid, p.23). They point out, for example, that differences in the use of subject pronouns (often dropped in Polish) and function words (more frequent in English) might affect the resulting diversity values. Bentz and Buttery point out another different language feature that can increase the lexical diversity of a particular language – the frequent use of borrowings (loanwords) (2014:40). Although it is evident that data lemmatization is able to perhaps reduce the effect of inflections on lexical diversity, the comparison of measurements from two or more languages can still prove to be problematic due to the simple fact that 'Languages display an astonishing diversity when it comes to lexical encoding of information' (ibid, p.42).

In the case of my study, I will be using VOCD measurements in order to compare the informants' use of two typologically different languages, Portuguese and English. Although it might be argued that the inflections in Portuguese would make such comparisons less feasible, it is useful to draw attention to what Xanthos and colleagues say about 'theoretical' and 'observed' morphological richness (2011:4). In their study of child speech and child-directed speech, they state that although a particular language may offer a potentially complex morphological system, 'As a rule, only a reduced fraction of the theoretical morphological richness of a system will be observed in any given sample' (ibid). With regards to *my* sample (the LOBILL Corpus), a simple frequency count of the present tense forms of the English verb 'do' ('do' and 'does') and the equivalent in Portuguese, 'fazer' ('faço', 'fazes', 'faz', 'fazemos', 'fazeis' and 'fazem') illustrates what these authors observed[25]. Despite there being potentially 6 present tense forms of the verb fazer, only the 'faz' form occurred while both English forms were found in the data. The fact that two of the main informants are children may also mean that more complex morphological features of the two languages are less in evidence in the corpus, reducing thus further the potential effect of these features on lexical diversity. Indeed, as will be seen in the discussion of the VOCD results (4.2), there seems to be little evidence

[25] Each form was searched for separately using the FREQ command (see item 7 in Table 5 for the search string used).

that the lexical diversity scores for Portuguese have been inflated by the language's theoretically rich morphology.

If the current study's sole focus were to compare lexical diversity measures across Portuguese and English, it would be a valid exercise to perform the VOCD analyses on both the raw data and on lemmatized data. By doing this the effect of inflections on the resulting D scores could be investigated. However, apart from practical issues relating to the process of lemmatization of data (a very time-consuming activity), my focus is on the investigation of *code-switching*, VOCD being used as a means to shed light on the relationship between languages participating in code-switched utterances. Before illustrating with example commands how I intend to use VOCD to this end, I will first discuss how the calculation of such diversity scores has the potential to contribute to research on code-switching, more specifically the relationship between the Matrix and Embedded Languages (see 2.1.1.1).

In the case of classic code-switching, where the Matrix Language (ML) provides the grammatical morphosyntactic structure into which mostly content words from the Embedded Language (EL) are inserted, I would like to hypothesize that greater lexical diversity (that is, higher D values) would be more evident in the EL and less so in the ML. This is because the repetitive use of a relatively small number of closed class items (articles, conjunctions, pronouns etc) would result in lower D scores for the ML when compared with the EL which typically draws from a much more varied pool of words. Thus the hypothesis would be that there is a relationship between D scores and the roles of the two languages in mixed utterances: a lower D score would be evidence that that particular language is more likely to be acting as the ML whereas a higher D score would indicate that the language in question is being used as the EL. The greater the difference between the two D scores (for English and Portuguese in this case), the closer the bilingual speech would approximate to classic code-switching. Where there is little or no difference between the two D scores one would be able to conclude that both languages are participating more equally in code-switching, sharing the grammatical structure and becoming more akin to what Muysken describes as congruent lexicalization (see 2.1.4.1). Having put forward this hypothsis, I will now very briefly discuss how I intend to use VOCD to test it.

As will be seen in 4.2 most of the analyses carried out with VOCD control for the effect of addressee. As discussed in the previous section, this is achieved via

KWAL (see the first part of the command lines below), the specified input then being sent to VOCD for analysis (the second part of the command lines):

(V1) **kwal @ +t%add +t"MEG" +s"PAI" +d +u | VOCD -s"@nonwords.cut" +r5 +s"[+ *]" -s"<@en>"**

(V2) **kwal @ +t%add +t"MEG" +s"PAI" +d +u | VOCD -s"@nonwords.cut" +r5 +s"[+ *]" -s"<@pt>"**

After KWAL has been used to select MEG's utterances addressed to PAI, VOCD is then instructed to analyse only code-switched utterances (+**s"[+ *]"**). By adding the string **-s"<@en>"** (see the first command line) the resulting D score will be for just the Portuguese tokens of the code-switched material. Likewise, with the addition of the string **-s"<@pt>"** (the second command line) the D score for only English tokens will be given in the output. These two D scores could then be analysed in the light of the hypothesis proposed above regarding the relationship between D scores and the ML/EL asymmetry of code-switched utterances.

Apart from taking into account the variable of addressee, by performing VOCD analyses on selected files (for example, only meal time interactions or recordings carried out over certain time periods), the variables of interaction type and time/age could also be examined. These findings could be triangulated with the results of other analyses (such as word lists and concordances) which may serve to provide further evidence for the proposed hypotheses. If it is found that there is a relationship between the D scores and the role of each language in bilingual utterances, this could lead to a new model which uses D scores to describe the particular nature of a bilingual's code-switching. In the form of a continuum, this could range from classic code-switching, where the D values for the participating languages are disparate to a fused lect or congruent lexicalization, where one would expect to find similar D values in each participating language.

Showing how to use such measurements of lexical diversity to describe a bilingual speaker's language use is one of the original methodological contributions of this study. A further contribution is discussed in the following section where I show how the CLAN command WDLEN can be exploited to provide further means for quantifying the differential roles of languages in code-switched speech.

3.3.6 WDLEN

Although the name indicates that the function of WDLEN (Word Length) is to calculate the length of words (in characters), it actually also calculates the length of utterances (in words). The output is displayed in the form of a table which shows the distribution of frequency of words and utterances in terms of their lengths. Mean Word Length (MWL) and Mean Utterance Length (MUL) are also provided. This command is particularly useful for researchers studying language acquisition and language development in children but could also be applied to the study of other areas such as second language learning. In this particular study I aim to use these two measurements in a novel way - as a method for establishing the relative participation of English and Portuguese in code-switched utterances in terms of the ML/EL Asymmetry.

Before I present my hypotheses regarding the relationship between mean word and utterance lengths and the ML/EL Asymmetry, it is important to briefly raise the issue of the comparability of word lengths across typologically different languages (such as English and Portuguese). Despite the diversity in how languages encode lexical information (see discussion in previous section), there are studies which demonstrate that when it comes to comparing the distribution and frequency of word lengths across many languages, similar patterns can be found. In Smith's 2012 study, for example, the distributions of lengths (in characters) of distinct words in the spell check dictionaries of 11 different languages were compared and found to be very similar (see Figure 2, Smith, 2012:12-13). Average word lengths also proved to be comparable, ranging from 8.3 characters in Swahili to 11.7 in German and, more importantly for the current study, the means for English and Portuguese proved to be very similar, 9.2 for the former and 9.9 for the latter. Such similarity had also been reported on in a previous study carried out by Piantadosi and colleagues (2011) in which they compared the relationship between word lengths (two-, three- and four-character words) and *frequency* and between word lengths and *information content* in 10 different languages. Again, the evidence proved to demonstrate that English and Portuguese are comparable, this time in terms of the frequency of shorter words (see Figure 1, Piantadosi et al, 2011:3527).

Although Smith's study provides support for the comparability of word lengths in English and Portuguese, the fact that he used *written* data (from dictionaries) for his analysis and performed calculations based on *distinct* words (i.e. types and not

tokens), means that it is unlikely that the word length values he found (above 9 characters for both languages) can be used as a baseline for the results of the analyses performed on the LOBILL Corpus (which contains naturally occurring spoken data). Evidently, when compared to written language, the economic and repetitive nature of spoken language will tend to lead to lower mean word length values in any language (as first pointed out by Zipf, 1935). However, what needs to be highlighted here is that both of the above-mentioned studies demonstrate that there are no fundamental differences in word length between English and Portuguese.

Having now discussed this issue of comparability, I will present my hypotheses regarding mean word and utterance lengths and the ML/EL Asymmetry. The two hypotheses proposed below are based on the differences between the typical contributions of the ML and EL in CS utterances. In terms of word length, we would expect that the grammatical nature of the ML would result in relatively low mean word lengths, pronouns, articles, auxilary verbs and so on being typically shorter in numbers of characters than lexical items. In contrast, by typically contributing lexically-laden words, we would expect to find relatively high mean word lengths for the EL. Based on these premises, my first hypothesis is that there is a relationship between mean word lengths and the role of a language in CS utterances: a low MWL would reflect the Matrix Language while a high MWL would be indicative of the Embedded Language.

The second difference between the typical contributions of the ML and EL in CS utterances is in terms of the quantity of words each language contributes to the utterance. In classic code-switching, a speaker will most frequently insert only single words or a small number of words from their EL in a CS utterance. This means that the remainder of the utterance consists of words belonging to the ML. Although WDLEN is normally used to calculate the mean utterance length of whole utterances, due to the language coding in the LOBILL Corpus I am able to split utterances in such a way that I can ask WDLEN to give me the MUL of the English part separately from the MUL of the Portuguese part. This means that I can test my second hypothesis which predicts that a low MUL (for the English or Portuguese part) will reflect the contribution of the Embedded Language while a high MUL will indicate a particular language acting as the Matrix Language.

Taking these two hypotheses together one could make the following predictions relating to the relative roles of the participating languages in CS utterances: a low mean word length coupled with a relatively high mean utterance length would indicate a language's role as the ML whereas a high mean word length coupled with a low mean utterance length would indicate that language's role as the EL. In data where the means prove to be less disparate, this would be an indication that a speaker's code-switching is less 'classic'. And where the means are very similar one would expect to find bilingual language use more akin to congruent lexicalization.

As is the case for VOCD, I am only able to exploit WDLEN in this novel way because of the language coding inserted in the corpus. Without such coding, it would be impossible to investigate the hypotheses proposed in this and the previous section.

In terms of what a typical WDLEN command line would look like, the following two examples show how the means (both word and utterance length) are achieved for each language:

(W1) **kwal @ +t%add +t"JAM" +s"PAI" +d +u | wdlen -s"@nonwords.cut" +s"[+ *]" -s"<@en>" +r5**

(W2) **kwal @ +t%add +t"JAM" +s"PAI" +d +u | wdlen -s"@nonwords.cut" +s"[+ *]" -s"<@pt>" +r5**

After using KWAL to select the specific speaker/addressee utterances I am interested in (the first part of the command lines), WDLEN is in then asked to perform its analysis on only the Portuguese material (**-s"<@en>"**) in CS utterances (**+s"[+ *]"**) (first command line) and then on only the English material in mixed utterances (the string **-s"<@pt>"** excluding any Portuguese material). As mentioned previously, the resulting output is in table format, providing not only the means but also the frequencies of each word and utterance length.

Such WDLEN analyses are particularly useful for highlighting any differences in how JAM and MEG use their two languages when code-switching. And the results from these analyses could also be compared to those of other bilingual speakers.

If it is shown in this study that the quantitative results from WDLEN and VOCD appear to correlate with the results from more qualitative analyses, it would be

feasible to develop a quantitively-based code-switching model which could be used to explain and predict CS patterns in other bilingual data which have been coded accordingly. This potential contribution to the field of code-switching research is discussed further in Chapter 8.

3.3.7 A summary of the switches and strings used to investigate code-switching in the LOBILL Corpus

In this section I provide a summary, in the form of a table, of the different types of switches and search strings used in the analysis of the LOBILL Corpus. This will serve as a useful reference for those readers particularly interested in the construction of the command lines. The switches, in column 2, have been placed in alphabetical order and column 3 shows how these switches are combined with different elements (numbers and codes). While column 4 shows the command(s) used with each particular search string (and switch), columm 5 informs us what that command is instructed to do. Where example combinations of switches and strings are given (column 3) this implies the possibility of substitutions of the target string. For example, in 3, the string +fJAM could be substituted by +fMEG or +FMOT etc. Similarly, in 7 through 10, the words 'going' and 'go*' could be replaced with any other words. Where substitution can occur this is indicated by the use of 'e.g' before the string.

Table 5. The switches and search strings used in the analysis of the LOBILL Corpus

		Combination of switch and string	**CLAN command**	**Instruction**
1	+d		KWAL	Removes extra information (such as file names and line numbers) from results so it can be sent ('piped') to a second analysis
2		+d1	KWAL	Includes line numbers in output
3	+f	e.g. +fJAM	ALL	Sends and saves output to a file with extension .cex
4	+o		FREQ	Puts results in order of frequency (and not alphabetically)
5	+r	+r1	WDLEN	Includes omitted material (characters) found in parentheses
6		+r5	FREQ VOCD WDLEN	Analyses original forms and not 'replacement' forms (found in square brackets)
7	+s	e.g. +s"going"	FREQ	Outputs frequency of 'going'

8			COMBO KWAL	Outputs utterances where 'going' occurs
9		e.g. +s”go*”	FREQ	Outputs frequency of variants of 'go' (such as 'go', 'goes', 'going', 'goed' etc)
10			COMBO KWAL	Outputs utterances where variants of 'go' occur
11		+s”[@en]”	FREQ	Outputs number of English language codes
12		+s”[@pt]”	FREQ	Outputs number of Portuguese language codes
13		+s”[@sp]”	KWAL	Outputs utterances containing Spanish codes
14		+s”<@sp>”	FREQ	Outputs list of Spanish tokens
15		+s”[+ *]”	FREQ	Outputs list of CS tokens
16			KWAL VOCD WDLEN	Selects all CS utterances for analysis
17		+s”[+ e*]”	KWAL	Selects all CS utterances beginning in English
18		+s”[+ p*]”	KWAL	Selects all CS utterances beginning in Portuguese
19		e.g +s”[+ epe]”	KWAL	Selects CS utterances which switch from English to Portuguese and back to English
20		+s”<+ *>”	FREQ	Outputs frequency list of CS codes
21		+s”[@tq]”	KWAL	Selects utterances in which tag questions occur
22			FREQ	Outputs number of tag question codes
23		+s”<@tq>”	FREQ	Outputs frequency list of tag question tokens
24		+s”[//*]”	KWAL	Selects utterances where retracings and reformulations occur
25			FREQ	Outputs number of retracing and reformulation codes
26		+s'[”]'	KWAL	Selects utterances containing metalinguistic language use
27			FREQ	Outputs number of metalinguistic codes
28		+s'<”>'	FREQ	Outputs frequency list of metalinguistic tokens
29		+s”[*]”	KWAL	Selects utterances in which errors occur
30			FREQ	Outputs number of error codes
31		+s”<*>”	FREQ	Outputs frequency list of error tokens
32		+s”*@mf”	FREQ	Outputs list of mixed form tokens
33			KWAL	Outputs utterances containing mixed form codes
34		+s”*@m”	KWAL	Selects utterances containing kinship variants for MOT
35			FREQ	Outputs frequency list of kinship variants for MOT

36		+s"*@p"	KWAL	Selects utterances containing kinship variants for PAI
37			FREQ	Outputs frequency list of kinship variants for MOT
38	-s	-s"<@en>"	FREQ VOCD WDLEN	Removes English tokens from the analysis
39		-s"<@pt>"	FREQ VOCD WDLEN	Removes Portuguese tokens from the analysis
40		-s"[+ *]"	KWAL VOCD WDLEN	Removes all CS utterances from the analysis
41		-s"@nonwords.cut"	FREQ VOCD WDLEN	Removes all non-words from the analysis
42	+t	e.g. +t*JAM	ALL	Selects utterances pertaining to a particular speaker, e.g. JAM
43		+t%add	KWAL	Outputs addressee tiers
44		e.g. +t%add +s"MOT"	KWAL	Selects utterances addressed to MOT
45		+t%err	KWAL	Outputs error tiers
46	-t*	-t* +t%add	FREQ	Outputs number of utterances addressed to each interlocutor

There is an important observation to be made about the use of the +s switch when searching for metalinguistic language use in the corpus. As can be seen in 26-28 above, it is necessary to enclose the search string (["]) in single quote marks rather than the usual double quotation marks. This, however, is the only exception to the rule. As for the consequence of using angled brackets instead of square brackets in the search strings with FREQ (see 14, 20, 23, 28, 31, 38 and 39), this has already been discussed in the section on FREQ.

The example command lines used to illustrate the functioning of the five CLAN commands showed how several different elements (switches and codes) can be combined to perform the desired analyses. The fact that the LOBILL Corpus contains an enriched level of coding means that the potential for analysis is greatly enhanced, as indicated in the discussions of each of the commands (COMBO, FREQ, KWAL, VOCD and WDLEN). However, it is in the discussion of the analyses and the results that we will see the full extent to which the LOBILL Corpus can be exploited in order to investigate code-switching.

The structure of the discussion over the following four chapters reflects my examination of the data as I progress from more quantitative analyses to a more qualitative interpretation of the bilingual informants' code-switching practices. The focus of Chapter 4 are the results of the purely quantitative analyses and I look at how such results can be interpreted in terms of the contribution both languages make to code-switched utterances. In Chapter 5, I discuss the results of word and code-level analyses, applying the 4-M model (see 2.1.1.1) to the word frequency data. In Chapter 6, I present a more qualitative analysis of the siblings' code-switched data as I examine the results of utterance-level analyses. In such an analysis I consider the effect of extra-linguistic factors on their code-switching practices. Whereas in Chapters 5 and 6 the data under analysis mainly come from the siblings' interactions with their parents, in Chapter 7 I briefly examine the code-switched utterances of other family speaker/interlocutor combinations (i.e. between the parents, between the siblings and when the parents address the siblings). Such examination allows for a more insightful interpretation of the code-switching practices of the siblings as I am able to consider aspects related to the language socialization occurring within the family unit.

4. Quantitative analyses and results

In the previous chapter the specific methodology used to investigate code-switching in the LOBILL Corpus was discussed in detail. The current chapter will now present the results of the quantitative analyses carried out using the CLAN commands and offer an interpretation of this data. I will begin by discussing the results obtained by using FREQ before moving on to analyse the results outputted by the commands VOCD and WDLEN. For readers who are particularly interested in the methodological aspects of using CLAN, the command lines for each analysis are included in the footnotes[26]. Information regarding the functioning of the different switches and strings which make up each command line can be found in Table 5 (found at the end of the previous chapter).

4.1 FREQ analyses and results

Although it is through more specific FREQ analyses that we will learn most about each speaker's language use, it is useful to first look at some general frequency data: this will provide us with an overview of the linguistic make-up of the LOBILL Corpus and its speakers.

4.1.1 General FREQ results for The LOBILL Corpus

A simple FREQ analysis[27] provides us with the total number of tokens (words) and types in the LOBILL Corpus: 137,227 tokens and 6473 types. By including the language codes in the search strings[28] this total is then broken down into the total number of words and tokens for each language: 107,351 tokens and 3826 types for English; 30,183 tokens and 2821 types for Portuguese[29]. By specifying code-

[26] Command lines are presented in TimesNew Roman. For batches of analyses where the only change in the command line involves the speaker code, only one example command line will be shown in the footnotes.

[27] freq @ +u +o -s"@nonwords.cut" +r5

[28] freq @ +u -s"<@pt>" +r5 +o -s"@nonwords.cut" and freq @ +u -s"<@en>" +r5 +o -s"@nonwords.cut"

[29] There is a discrepancy of an additional 307 words if one adds the separate language totals together (107, 351 + 30,183 = 137,534) and compares this total with the overall token/type count (137,227). This is also the case for the word type total where the discrepancy is an additional 174 types when the separate language totals are added together (3826 + 2821 = 6647) and compared to the word type frequency count for the corpus as a whole (6473). As my language exclusion method (using -s) would mean that items coded as mixed forms (@mf) and Spanish (@sp) could potentially appear in both language lists (i.e. counting twice), I experimented excluding these from the analyses shown in footnotes 27 and 28 by adding the strings -s"*@mf" and -s"*@sp". This, however, only accounted for 15 of the additional tokens and 9 of the word types. As the purpose of these analyses was to give a general overview of the LOBILL Corpus I decided to expend no more

switched utterances[30] we are then able to learn what proportion of the totals is made up of code-switched discourse; 9407 tokens and 1831 types. The two graphs below summarise these totals:

Figure 3. Total number of tokens in the LOBILL Corpus per language mode

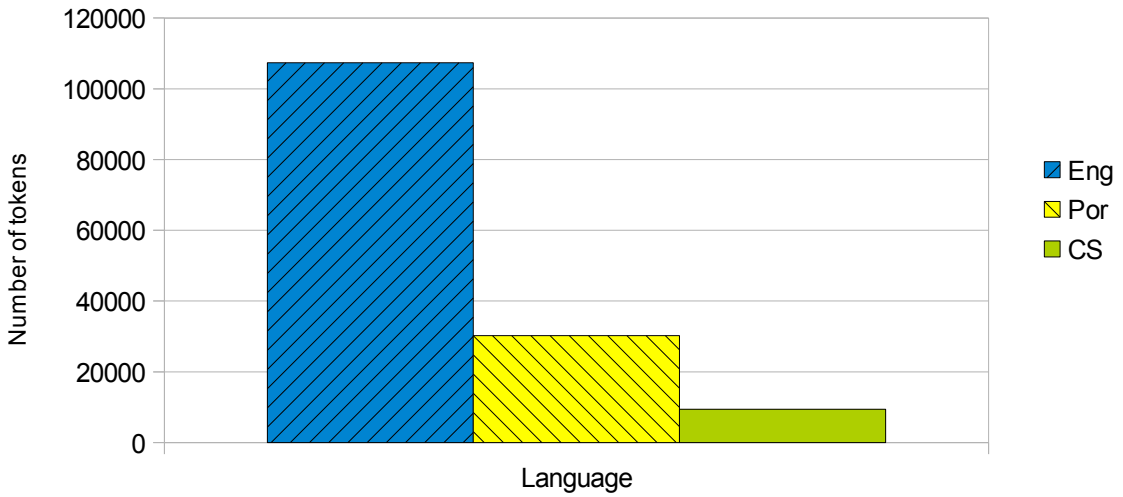
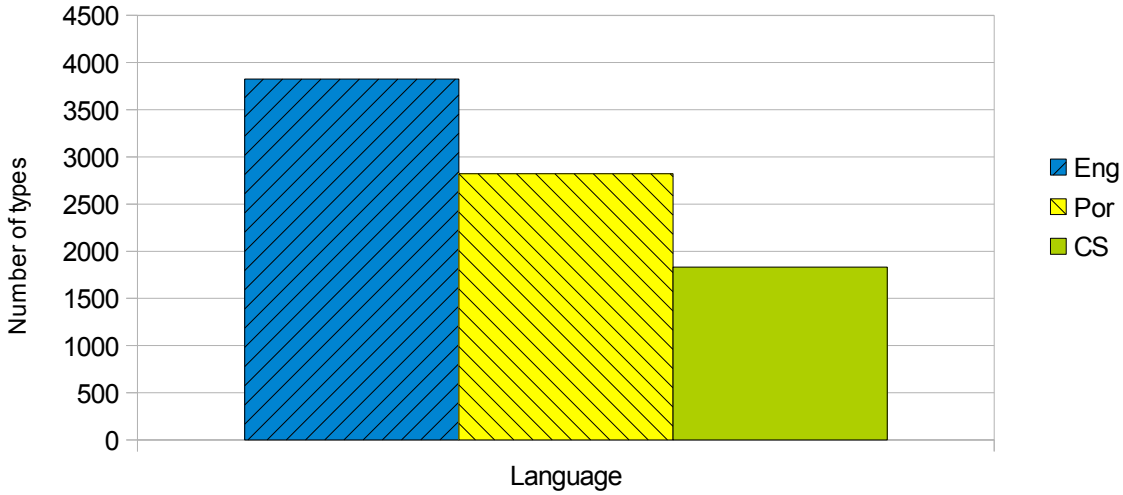


Figure 4. Total number of types in the LOBILL Corpus per language mode.



time investigating these discrepancies which I believed would have relatively little impact in terms of the frequency output of individual speakers.

[30] freq @ +u +s"[+ *]" +o -s"@nonwords.cut" +r5

The data in Fig. 3 shows us that English is the most spoken language (in terms of numbers of words) in the corpus, accounting for over two thirds of the overall total. While Portuguese tokens account for under a third of the total, CS tokens make up less than a tenth of the total. A similar pattern is found in Fig. 4, which shows the totals of types for each language, although the differences are much less marked. Indeed, it is important to highlight even at this stage that considering the relatively small number of *tokens* contributed by CS discourse, the number of CS *types* appears to be disproportionately high compared to the totals for English and Portuguese. Of course, the fact that the type/token ratio will naturally become smaller and smaller as the total number of tokens increases (see discussion in section 3.3.5) might be one way of explaining why the 3,800 types for English appears to be comparatively low. However, as will be seen in section 4.2, by using vocabulary diversity measures it is possible to ascertain that the high number of types for CS discourse found in the LOBILL Corpus *is* actually significant and reflects a particular feature of code-switched utterances.

Having given a general overview of the linguistic make-up of the LOBILL Corpus, it is now useful to break down the overall totals shown above into speaker totals i.e. the number of tokens and types each speaker contributes to the corpus.

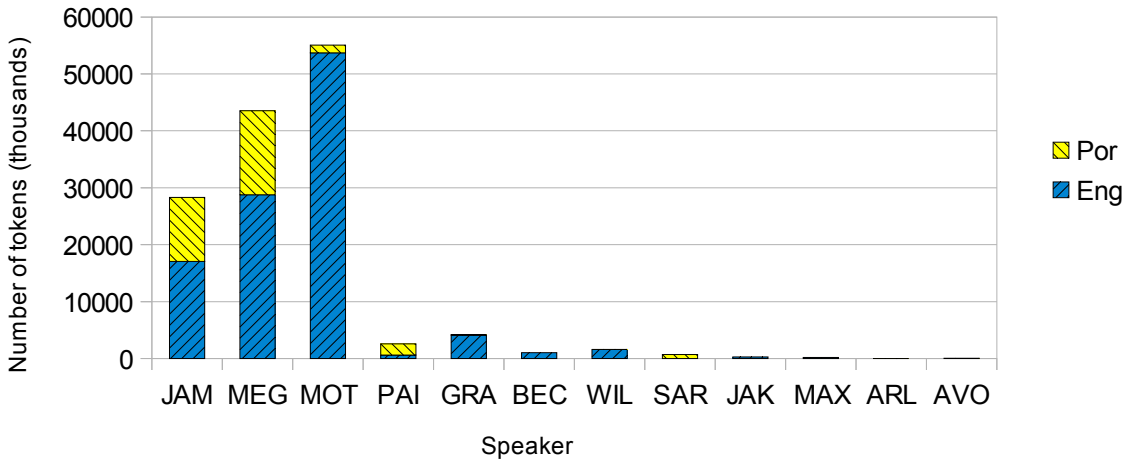
4.1.2 FREQ results per speaker

By simply specifying the speaker code in the FREQ command line[31], it is possible to obtain total tokens for each of the 19 speakers who feature in the LOBILL Corpus. For seven of these speakers (DAN, GRD, JAN, JUL, ROS, VIN and VOV) the output of the frequency lists was zero. The reason for this is that due to their singular role as interlocutors in the telephone conversations with the siblings their speech was not recorded. Their perceived turns are transcribed with the symbol **www** which is ignored by FREQ. Although there may not be any output for these seven speakers, their roles as monolingual interlocutors will allow us to factor in the crucially important addresse variable when it comes to analysing the code-switching behaviour of the two main informants.

For each of the twelve remaining speakers, two further analyses were carried out[32] in order to determine the number of tokens of English and Portuguese words which constituted their overall totals. This relative frequency can be seen in Fig. 5.

[31] freq @ +t*JAM +u +o -s"@nonwords.cut" +r5

Figure 5. Totals of tokens for English and Portuguese per speaker in the LOBILL Corpus



Again, although the output for nine of the speakers (PAI, GRA, BEC, WIL, SAR, JAK, MAX, ARL and AVO) appears to be rather insignificant in terms of totals, the frequency lists did reveal that apart from the siblings' father (PAI), the other eight speakers produced only monolingual utterances: the British relatives (GRA, BEC, WIL, JAK and Max) only used English and the Brazilian participants (SAR, ARL, AVO) used Portuguese exclusively. The analysis of the siblings' interactions with these monolingual speakers will allow us to see if and how the variable of an addressee's monolingualism affects the children's code-switching practices.

It is clearly evident from the chart that the three main informants (JAM, MEG and MOT) produce the majority of the tokens in the LOBILL Corpus: combined, their totals account for over 90% of the overall total. It is not suprising that MOT's total is higher than JAM's or MEG's: she was present in almost every interaction and her role as an adult caregiver would presumably mean her contribution in terms of tokens is likely to be higher than a child's. It is also of little surprise that MEG's contribution surpasses that of her younger brother as she is almost two and a half years older and therefore more developed linguistically. What *does* appear to be worthy of note from the chart is the relative proportion of English and Portuguese tokens for these three speakers. Although they use both English and Portuguese, the latter accounts for over a third of all tokens for both JAM and MEG, their respective Portuguese

[32] freq @ +t*JAM +u -s"<@pt>" +r5 +o -s"@nonwords.cut" and freq @ +t*JAM +u -s"<@en>" +r5 +o -s"@nonwords.cut"

token counts being 11,262 and 14,775. In contrast, MOT's Portuguese token count (1,364) represents less than a tenth of her total token count. This seems to indicate that while MOT uses English almost exclusively, her children have recourse to Portuguese much more frequently. From the data in Fig. 5 it is not possible to determine whether these Portuguese tokens are found in monolingual utterances or are used in code-switched utterances. However, we know from Fig. 3 that almost 7% of the total number of tokens in the LOBILL Corpus do indeed occur in code-switched utterances. In order to investigate this further, the utterances of only the first four (bilingual) speakers from the chart (JAM, MEG, MOT and PAI) were analysed again using FREQ. As the remainder of the speakers only produced tokens in one language (either English or Portuguese), this meant that code-switching could not have occurred in their discourse and further analyses of their utterances with FREQ were not necessary. The next section discusses what these more specific FREQ analyses revealed about the four bilingual speakers' language use.

4.1.3 FREQ results for the code-switched utterances of the siblings and their parents.

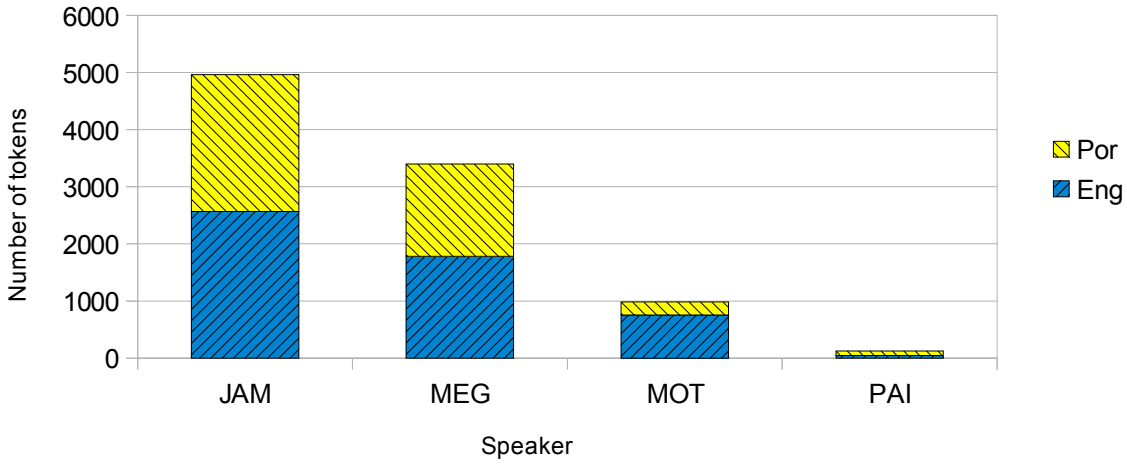
Three analyses were carried out on the code-switched utterances of the four main informants to obtain (i) the total number of tokens for their code-switched speech[33] and the number of (ii) English tokens[34] and (iii) Portuguese tokens[35] which make up these totals. The results of these three analyses can be seen in Fig. 6.

[33] freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" +r5

[34] freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" -s"<@pt>" +r5

[35] freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" -s"<@en>" +r5

Figure 6. Totals of English and Portuguese tokens in CS utterances per bilingual speaker



Immediately evident from the chart is the contrast that can be seen between the four bilingual speakers in terms of how much code-switching they engage in. JAM's token count for his code-switched utterances totals almost five thousand (4,925), representing approximately 17% of his overall token count (28,207). This contrasts with MEG's CS total (3,376 tokens) which only represents approximately 8% of her overall token count (43,428). The total number of CS tokens for MOT (981) is very low accounting for a mere 1.7% of her overall total (55,017) and PAI's CS total is a mere 122 tokens which represents 4.7% of his low overall total of 2,555 tokens. It is clear that JAM and MEG appear to engage in code-switching to a greater extent than their mother, and that JAM appears to code-switch significantly more than his sister (given their relative overall contributions to the LOBILL Corpus in terms of tokens).

If we now look at the proportion of English and Portuguese tokens which make up each of the four speakers' code-switched utterances, we find significant similarities and differences. Firstly, despite the difference in the totals of overall CS tokens, the siblings share exactly the same proportions of English and Portuguese tokens: the former accounts for 52% of the total while the latter accounts for 48%. Is this evidence that JAM and MEG use their two languages in the same way? This cannot be ascertained here but is an indication of what will be revealed by more qualitative analyses in Chapter 6. Of MOT's total CS tokens 77% are English and 23% are Portuguese, suggesting that her maternal language appears to play a more

dominant role when she code-switches. By the same notion, despite the very low number of tokens, PAI appears to favour his native Portuguese (63%) more than English (37%) when engaging in code-switching. From the data in Fig. 6 it might be tempting to assume that both JAM and MEG use English and Portuguese in roughly equal proportions in their code-switched utterances i.e. in an utterance of eight words, four would come from each language. However, it will only be possible to confirm or refute this by more qualitative analyses of the data.

The data in Fig. 6, in conjunction with that in Fig. 5, have allowed us to make important primary observations about the four bilingual speakers and their language use. However, as will be seen in the following discussion, there is a variable that is of upmost importance to take into account if we are to interpret the data correctly – that of addressee.

4.1.4 FREQ results for the code-switched utterances exchanged between the four family members

In the discussion so far we have established that only four of the speakers of the LOBILL Corpus use both English and Portuguese in their discourse. Although there is no evidence that the seven telephone interlocutors (whose turns were not recorded) only used their maternal tongue when speaking to the siblings, their status as monolingual language users supports the assumption that bilingual language use was restricted to the four main informants. In the last section we also learnt that the siblings and their parents all engaged in code-switching but to varying degrees. The analyses in this section will take into account the variable of addressee and the results will reveal any relationships between the speakers' use of English and Portuguese in CS utterances and their addressees.

In order to carry out these analyses, search strings specifying a particular addressee were systematically included in each command line. For example, having already asked CLAN to only select JAM's utterances, I then simply substituted the addressee code each time until I had performed the same analyses on JAM's utterances for all 18 addressees. For each specified addresse a batch of six analyses were performed, the output providing (i) the number of overall tokens addressed to that interlocutor[36], (ii) the number of English tokens[37], (iii) the number of Portuguese

[36] kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o -s"@nonwords.cut" +r5

[37] kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o -s"<@pt>" +r5 -s"@nonwords.cut"

tokens[38], (iv) the number of CS tokens[39], (v) the number of English tokens in only CS utterances[40], and (vi) the number of Portuguese tokens in only CS utterances[41]. Despite this resulting in a potential 432 analyses (4 speakers x 6 analyses x 18 addressees), a zero frequency output for (i) or at (iv) would mean that no subsequent analyses for that particular addressee would be necessary. For example, if no CS tokens were found in JAM's output when addressing his uncle WIL (analysis (iv)), then it follows that analyses (v) and (vi) would also result in zero frequency (therefore making them unnecessary).

Before discussing the results of these analyses, it is important to make the following observation. Despite the fact that 15 of the speakers/interlocutors in the LOBILL Corpus were classed as monolinguals and that there was no evidence of their using a second language, it would be wrong to assume that the bilingual speakers did not use their 'other' language with these monolingual interlocutors. Indeed, if the frequency results were to flag up code-switched tokens in the word lists of the siblings or their parents when addressing monolingual interlocutors this would be significant and require further qualitative analysis.

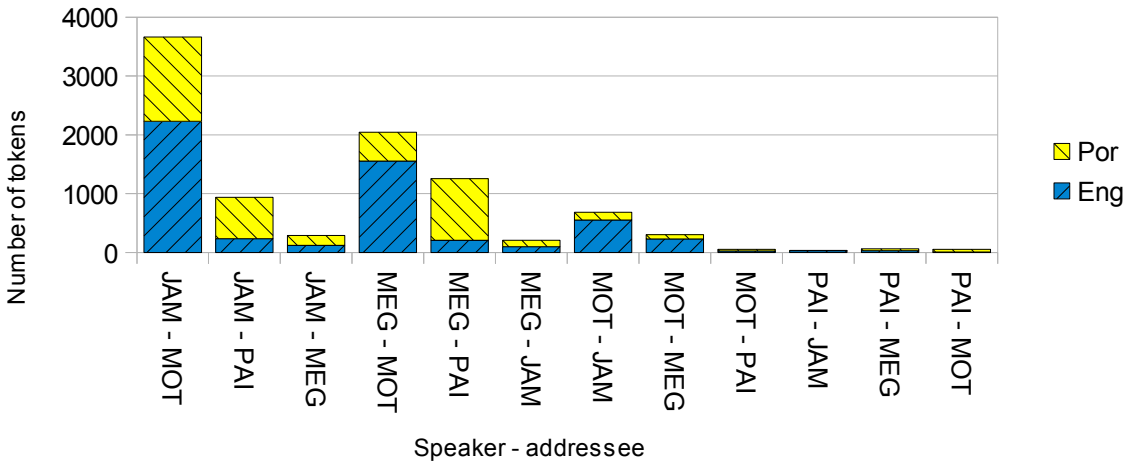
As the focus of this study is on the code-switched speech of the speakers, the chart below (Fig. 7) shows the results for analyses (iv), (v) and (vi), revealing the total tokens and the proportion of English and Portuguese tokens in only CS speech. The results of analyses (i), (ii) and (iii) will be used in the discussion as a means of making comparisons and relativising the data. The fact that only the results of the same three addressees per speaker are shown on the chart does not mean that CS tokens were not found in the utterances addressed to the other 15 interlocutors. However, as the numbers were so low, their inclusion in the chart would have made them visually imperceptible. Only a more qualitative approach to these tokens will enable the investigation of their potential significance in the data (see the utterance-level analyses in Chapter 6).

What strikes the reader on first examination of the data in the chart are the differences that can be seen between the overall totals of CS tokens of the different speaker/interlocutor combinations. For example, JAM appears to engage in far more

[38] kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o -s"<@en>" +r5 -s"@nonwords.cut"
[39] kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" +r5
[40] kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5
[41] kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@en>" +r5

code-switching with MOT than with PAI or MEG. A similar pattern, albeit less extreme, can be seen in the total CS tokens MEG addresses to MOT, PAI and JAM.

Figure 7. Number of English and Portuguese tokens in CS utterances per addressee.



It is crucial to see what these CS totals represent in terms of proportions of the overall number of tokens addressed to each interlocutor (analysis (i)). When this is done we find that first impressions turn out to be misleading. For example, JAM's CS tokens to MOT (3,655) account for 18% of the total tokens (19,592), those addressed to PAI (920) make up 25% of the total (3,670) whereas those CS tokens directed to MEG (280) amount to approximately 9% of the total tokens (3,208). Therefore, it would appear that there is evidence that JAM actually engages in more code-switching with his father (PAI) than with his mother (MOT) and significantly less with his sister (MEG). This is also true of MEG's data: of the 28,989 tokens she addresses to MOT, 7% (2,039) are CS tokens; of the 6,657 tokens adddressed to PAI, 19% (1,251) are CS tokens; and of the 6134 tokens she addresses to JAM only 3% (197) are CS tokens. As for JAM, MEG appears to code-switch more with her father than her mother and much less with her brother. Although the patterns for the siblings are similar, a comparison of the percentages shows that JAM code-switches relatively more than MEG when interacting with the same interlocutors. Only qualitative analyses will shed light on possible explanations for these differences.

If we now look at what the chart tells us about the CS tokens for speakers MOT and PAI, what we find are very low overall totals, the highest being the 682 CS tokens addressed by MOT to her son JAM. Although PAI's CS totals for each interlocutor are very low (37 to JAM, 61 to MEG and 53 to MOT), as percentages of the overall total number of tokens addressed to each interlocutor (4%, 5% and 4%), they are actually higher than those of MOT. The latter's CS totals when addressing each interlocutor (682 for JAM, 304 for MEG and 53 for PAI) account for only 2%, 1% and 3%, respectively, of the overall token totals (29,839, 24,660 and 1762). Such low numbers of tokens and percentages indicate that both MOT and PAI's use of code-switching with their children and with each other is very limited. In terms of language socialization, if both parents engage very little in code-switching, we might expect to find similar patterns in the speech of their children. However, we find data suggesting that the siblings code-switch significantly more than their parents. Why is this? When the discussion progresses from quantitative to more qualitative results, explanations for these differences will emerge.

The discussion above was concerned with the overall number of CS tokens of each bilingual speaker per addressee. Now we will examine what the chart tells us about the proportion of English and Portuguese tokens which make up these totals. Beginning with speaker JAM, the data reveals that when code-switching with his mother (MOT) English plays a greater role (in terms of number of tokens) than Portuguese: 61% (2243 tokens) are English tokens while 39% (1435) are Portuguese tokens. When addressing his father (PAI) the opposite is true: 25% (233) are English tokens and 75% (702) are Portuguese tokens. A similar pattern is found for MEG when addressing the same two interlocutors: 76% (1560) of the CS tokens directed at MOT are English, the remaining 24% (494) being made up of Portuguese tokens; and with PAI as addressee we find a reversal of proportions, 16% (204) of English tokens and 84% (1050) of Portuguese tokens. This evidence suggests that for both JAM and MEG, English plays a more dominant role than Portuguese when they code-switch with their mother while with their father Portuguese is clearly more prominent. This appears to be indicative of the Matrix/Embedded Language asymmetry which characterizes classic code-switching. Moreover, the fact that the proportions of English and Portuguese tokens for MEG are more disparate than those for JAM may mean that MEG's code-switching is more 'classic' than that of her brother. One question to ask is why JAM has relatively more recourse to Portuguese

when code-switching with his mother and makes more use of English than MEG when code-switching with his father. Such a question will be answered as we delve further into the data and examine the effect of different variables on the siblings' code-switching practices.

When we examine the CS tokens the siblings use with each other we find percentages that do not appear to reflect asymmetrical use of English and Portuguese. When code-switching with his sister JAM's use of English accounts for 43% (123) of the total tokens and Portuguese accounts for 57% (164) of the total. MEG's percentages are very similar: 47% (99) for English and 53% (106) for Portuguese. Does this mean that both languages are participating more equally in code-switched utterances and that there is no Matrix or Embedded Language? It would be unwise to make such assumptions as these percentages may simply reflect an averaging of the proportions. For example, two utterances (of 10 words each) with equal proportions of English and Portuguese tokens (5 of each) would average the same (10) as two utterances (of 10 words each) where one contained a higher proportion of English tokens (7) to Portuguese tokens (3) and the other contained more Portuguese tokens (7) than English tokens (3). In both cases, the resulting proportion would be 50% for each language. Therefore, it might be that the FREQ output is neutralizing any potential difference in Matrix/Embedded Language use that exists in the siblings' code-switched utterances. However, if this were the case, it would mean that the siblings were not consistent in their code-switching patterns with each other, constantly changing their Matrix Language from utterance to utterance. Although this is certainly possible, it is more common for bilingual speakers to maintain the same Matrix Language with the same addressee. Simply looking at percentages in this case is not sufficient and again highlights the need to combine quantitative analyses with a qualitative examination of the data (see 7.2).

Before leaving this discussion of the chart in Fig. 7, we will take a brief look at the proportion of English and Portuguese tokens in the CS utterances of the siblings' mother. Whereas the CS totals for PAI (37 when addressing JAM, 61 to MEG and 53 to MOT) are too low to warrant further quantitative discussion, the higher CS totals addressed by MOT to the siblings enable us to examine the role each language might play in terms of the Matrix/Embedded asymmetry. Although the total number of CS tokens MOT uses with JAM (682) is double that which she uses with MEG (304), the proportion of English to Portuguese tokens is comparable: 80%/20% (JAM) and

75%/25% (MEG). These percentages reveal that the mother's CS utterances contain substantially more English tokens than Portuguese tokens. Even taking into account the potential problem presented by frequency averaging across utterances, it does seem likely that English plays a more dominant role in MOT's CS utterances directed to her children. Whether this role is the typical one of the Matrix Language will be revealed by further analyses. With regards to the CS tokens addressed to PAI, the total (53) is again too low to offer any interpretation. With only 24 English tokens and 29 Portuguese tokens all that can be said at this point is that there is indeed evidence of MOT engaging in code-switching with her husband.

At the beginning of this discussion it was pointed out that the results of the FREQ analyses for the remaining 15 addressees were insufficient to be included in Fig. 7. However, despite the very low number of tokens, it is important to highlight here that three of the speakers, JAM, MEG and MOT, did code-switch with some of these addressees. The raw figures for these results are shown in the table below and merit brief comment.

Table 6. Total number of CS tokens, English CS tokens and Portuguese CS tokens addressed by JAM, MEG and MOT to other interlocutors.

Speaker	Addressee	Language of addressee	Total CS tokens	English CS tokens	Portuguese CS tokens
JAM	GRA	Eng	122	70	52
	BEC	Eng	5	3	2
	AVO	Por	115	10	105
	VOV	Por	11	2	9
	VIN	Por	45	7	38
MEG	GRA	Eng	6	5	1
	BEC	Eng	2	1	1
	VOV	Por	24	4	20
	SAR	Por	15	2	13
MOT	MAX	Eng	6	5	1
	JAK	Eng	6	5	1
	VIN	Por	7	1	6
	SAR	Por	4	1	3

With such low figures one might think any attempt at an interpretation would be fruitless. However, if comparisons are made between the proportions of English and

Portuguese words directed to the addressees we do indeed find an indication of a relationship between the main language used by the bilingual speaker in these CS utterances and the addressee's mother tongue: where the addressee is a monolingual Portuguese speaker (AVO, VIN, VOV and SAR) all three speakers use correspondingly more Portuguese tokens than English tokens; where the addressee is a monolingual English speaker (GRA, BEC, MAX, and JAK), English tokens are correspondingly more frequent than Portuguese tokens. This again appears to support the existence of a Matrix/Embedded Language asymmetry at work in these code-switched utterances. However, there is one particular case where the numbers are rather unexpected. The table shows that JAM code-switches with his English Grandma (GRA) for a total of 122 tokens. Taking into account GRA's monolingualism it is rather surprising that 52 of these tokens are actually Portuguese! This unexpected output will be examined at utterance level in Chapter 6. For all of the other addressees the number of tokens in the 'other' language is comparatively low. The fact that we find evidence, however little, of JAM, MEG and MOT's use of code-switching with 'monolingual' speakers is intriguing and it is only by examining the concordances themselves (in Chapter 6) that explanations can be sought for such linguistic behaviour.

The discussion in this last section has focussed on the variable of addressee as a factor affecting the roles of languages (in terms of numbers of tokens) in the code-switched utterances of our four bilingual speakers. We have found evidence for the existence of a Matrix/Embedded Language asymmetry which needs further investigation in order to determine exactly how this asymmetry is realized in the speech of the four main informants. However, before the data is approached from this more qualitative angle (in Chapters 5 and 6), first the results of two other types of quantitative analyses, carried out with the CLAN commands VOCD and WDLEN, will be discussed.

4.2 VOCD analyses and results

As mentioned in 3.3.5, vocabulary diversity measurements are commonly used in the field of child language development and also in research in second language learning. At the time of writing it is believed that such a measure has not previously been used in order to investigate code-switching in electronic corpora. The discussion that follows, therefore, will be of particular interest to those who wish to

exploit existing measures in novel ways. The section will begin with a brief mention of what the Diversity (D) scores can tell us more generally about each speaker's language output. Then, the variable of addressee will be incorporated into the analyses and potential relationships between D scores and the Matrix/Embedded Language asymmetry will be discussed. The last part of the section will focus on examining the possible relationships between D scores and other variables, such as type of interaction and time periods. As for the previous section on FREQ, the command lines used in the analyses will be specified in the footnotes. And as for the results themselves, all D scores have been rounded to the nearest whole number[42]. This has been done to enable clearer presentation of the data.

4.2.1 General VOCD analyses and results

In the discussion of the FREQ results it became increasingly evident that it is only by taking into account the variable of addressee that we can begin to interpret with more accuracy the resulting numbers and proportions and what they represent in terms of Matrix/Embedded Languages. This is also true of D scores and, as such, most of the VOCD analyses systematically incorporated addressees in the command lines. However, a brief look at some more general results (i.e. non-addressee specific) did prove to reveal something interesting about the difference between monolingual and bilingual (code-switched) utterances in terms of D scores. These scores also provided a useful baseline for later comparison when the analyses became more specific.

First, a simple VOCD analysis[43] was performed on each of the 12 speakers (eight monolingual and four bilingual)[44] providing an overall D score which, in the case of the bilingual speakers, did not differentiate for language. Then by specifying the language in the command line, separate D scores for English and Portuguese were calculated for the four bilingual speakers[45]. Finally, one further VOCD analysis

---

[42] Note that as the VOCD programme performs its analysis on a random selection of the data, replication of the command lines shown in the following footnotes may result in slightly different D scores to the ones discussed in this section.

[43] kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut"

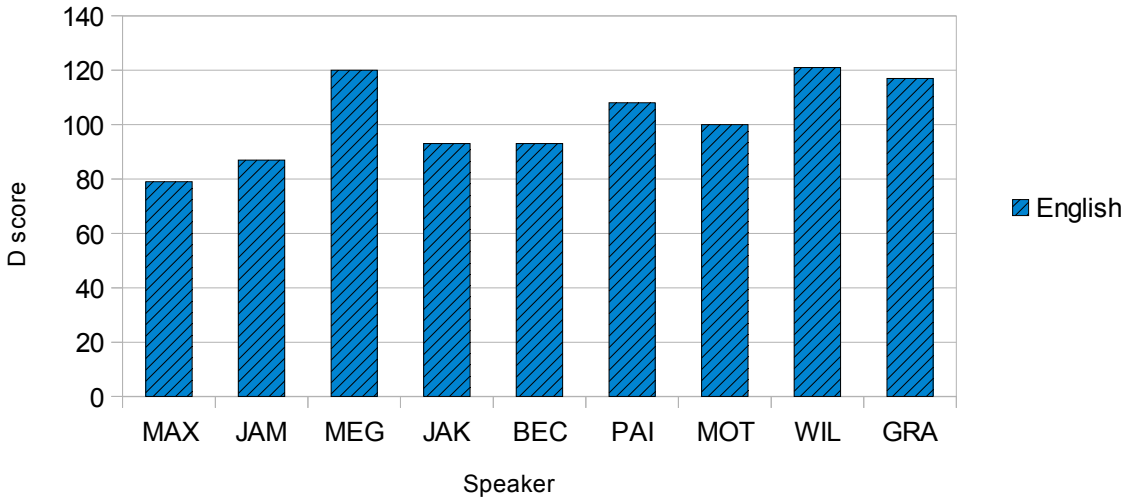
[44] Recall that the remaining seven monolingual informants were telephone interlocutors only and their turns were not recorded, therefore resulting in zero token output.

[45] kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut" -s"<@pt>" and kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut" -s"<@en>"

on only code-switched utterances provided D scores for the code-switched material[46] produced by JAM, MEG, MOT and PAI.

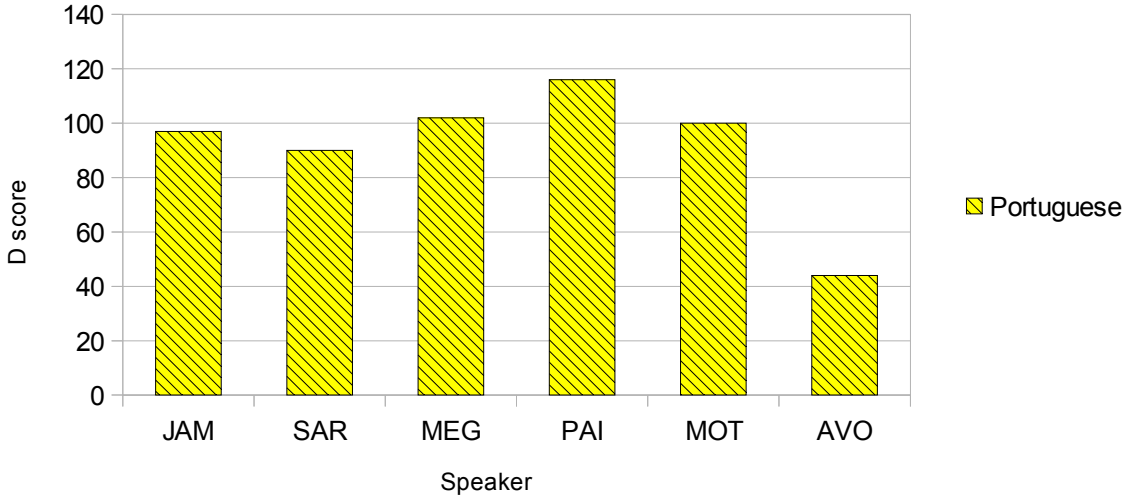
The results in the following two graphs (Figs. 8 and 9) show the D scores for eleven of the twelve speakers in the LOBILL Corpus. As one of the speakers, ARL, produced less than 50 tokens (the minimum number of tokens VOCD needs in order to output a D score), there was no result for her. The first graph shows the D scores for English output while the second shows the D scores for Portuguese material. The speakers have been ordered in terms of age, from youngest to oldest (the children being MAX, JAM, SAR, MEG and JAK). It is important to recall, particularly in the case of the child speakers, that the data in the LOBILL Corpus spans three and a half years and, as such, for several of the speakers the general D scores shown in the charts are not representative of a particular point in time but rather of an averaging of vocabulary diversity over this time span. The effect of this variable of time (and age) on D scores will be examined in section 4.2.4.

Figure 8. D scores for English tokens per speaker.



[46] kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]"

Figure 9. D scores for Portuguese tokens per speaker.



Here I am interested in making general comparisons of the D scores across the two languages and what we find is that for both English and Portuguese the majority of the speakers (monolingual and bilingual) have D scores between 80 and 120. The lowest score for English is 79 (belonging to MAX who is the youngest child) and for Portuguese we have a low 45 (belonging to AVO, the siblings' Brazilian grandmother)[47]. If we remove the low D score for AVO, and then calculate the average D score across speakers separately for English and Portuguese we arrive at a D score of 101.7 for English and 101.6 for Portuguese. An Independent Samples T-test confirmed that there was no significant difference in D scores between the two languages (t=0.23, df=12, p=.982) and this finding means that I will be able to use a rounded up D score of 102 as a baseline for the interpretation of more specific results when relationships are made between a language's D score and its role in terms of the Matrix/Embedded Language asymmetry in code-switched utterances.

If we now focus on the D scores of our bilingual speakers (JAM, MEG, MOT and PAI), it appears that their scores for each language are comparable to those of the monolingual speakers. With regards to JAM and MEG, this finding is interesting if we consider that it is commonly held that a bilingual child's vocabulary performance in each language falls below that of a monolingual child of the same age (Gathercole

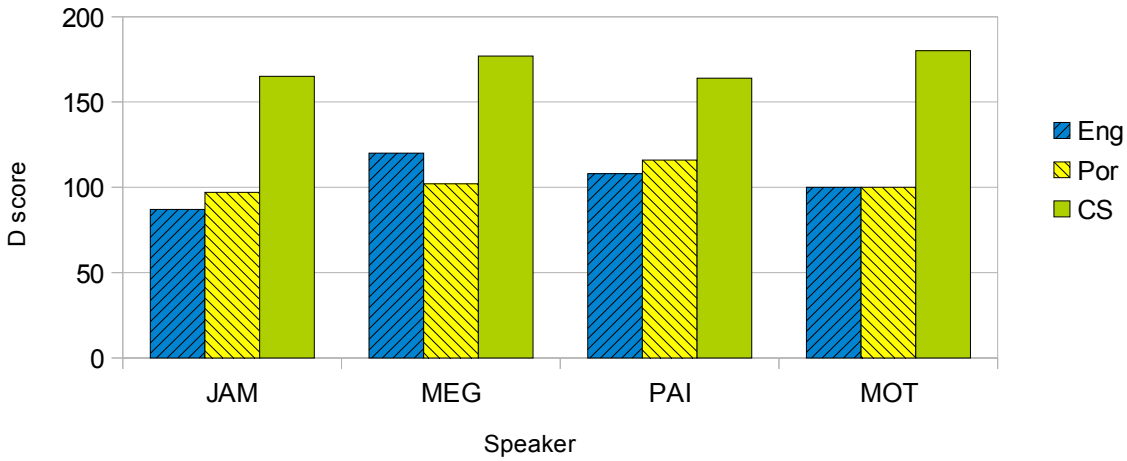
[47] A quick KWAL analysis of AVO's utterances (kwal @ +t*AVO +u +d) revealed that her total of 64 tokens and 44 types distributed over 12 utterances all occurred in one interaction in which she instructs her grandchildren on how to make plaster of paris figurines. Her repetition of the top 6 most frequent words accounts for 27 of the tokens and explains why her D score is comparatively low.

et al: 2008; Hoff et al: 2012;). In most of these studies a bilingual child's 'performance' is usually assessed by using standardized assessment measures based on *receptive* knowledge. Although researchers may recognize the need to take into account variables such as language exposure (at home and in the community) in order to refine assessment measures of bilingual children (Gathercole et al: 2013), it is unlikely that the results of receptive normed tests will ever accurately represent each bilingual child's knowledge of his or her vocabulary (and grammar). Clearly, an examination of a child's *productive* output would be much more insightful but practical implications in terms of data collection (especially the time expended on transcribing the data) would make this approach less viable.

Returning to the D scores of the bilingual speakers, it is evident that their scores for each language do not fall below those of their monolingual peers. In fact, the results of the first VOCD analysis (which did not differentiate for language) reveal their overall scores to be rather high when compared to their monolingual peers: the overall D scores for JAM and MEG were 170 and 183 respectively! Although it is beyond the scope of this study to elaborate on the implications of such findings it is worth pointing out the potential of such a measurement in contributing to further research on vocabulary development in bilinguals when compared to monolinguals.

To account for the particularly high overall D scores for JAM and MEG mentioned above, we will now turn to the results of the fourth VOCD analysis which provided D scores for the code-switched material of each bilingual speaker. To aid comparison the CS results are presented in Fig. 10 alongside those already shown in the charts above (note the change in scale).

Figure 10. D scores for English, Portuguese and CS tokens of the bilingual speakers.



What is immediately evident from the chart is that the D scores for the code-switched material for each bilingual speaker are markedly higher than those for either English or Portuguese. All four speakers score between 160 and 180 for CS material as opposed to a maximum of 120 (MEG) for any single language. With such a consistent pattern I believe it is possible to ascertain that these scores are reflecting the lexical richness of the code-switched discourse in the LOBILL Corpus. Whether such results would be found if the same measures were used on other bilingual language corpora remains to be seen but it is evident that the use of a quantitative measurement like D scores opens up the possibility of making comparisons between monolingual and bilingual speakers within the same corpus and ultimately across different types of corpora (monolingual and bilingual). As long as the methodology is replicable (and this study aims to enable this replicability), such comparisons would be a valid way of investigating differences in vocabulary diversity among different groups of speakers. This could lead to potential practical applications in areas such as child bilingual language development and education as well as second language research.

For the purposes of this study, what these general D scores have shown so far is that our bilingual speakers are comparable to the monolingual speakers in terms of vocabulary diversity in each language: the combined average D scores of the four bilinguals for English (103) and Portuguese (104) are just above the overall average score of 102. However, the results have also revealed that the code-

switched utterances of the LOBILL Corpus are characterized as being particularly lexically rich. Especially in the case of JAM and MEG this has meant that their overall D scores of 170 and 183 far surpass the established average of 102.

This brief discussion of the general VOCD results has served to establish baselines which will be useful for the analysis of more specific VOCD results. It has also served to highlight the apparent lexial diversity that can be found in CS speech. What the following section sets out to do is investigate what D scores can tell us about the role that each language plays, in terms of the Matrix/Embedded Asymmetry, in each bilingual speaker's code-switched speech. As we have already learnt, a crucial factor affecting these roles is that of addressee and therefore, unlike the analyses in the section above, all of the following analyses are addressee specific.

4.2.2 VOCD analyses and results of the code-switched utterances exchanged between the four family members

Since the focus of this section is on examining the VOCD results of only code-switched speech, it is evident that it would not be necessary to perform such analyses on the speech of the monolingual informants. For the four bilingual informants, frequency results (reported on in 4.1.4) had shown very low totals of CS tokens for certain speaker/interlocutor combinations. Although VOCD would have been able to provide an overall D score (i.e. English and Portuguese tokens combined) for some of these combinations, VOCD would not have then been able to provide separate D scores for each language (a minimum of 50 tokens is needed). As my interest is in analysing the relationship between the ML/EL asymmetry and the separate D scores for each participating language, it only makes sense to perform VOCD analyses on those speaker/interlocutor combinations where separate D scores can be calculated. Based on the frequency results, we are therefore left with eight speaker/interlocutor combinations on which the VOCD analyses can be carried out effectively: JAM/MOT, JAM/PAI, JAM/MEG, MEG/MOT, MEG/PAI, MEG/JAM, MOT/JAM and MOT/MEG.

Three VOCD analyses were performed on each of the above eight combinations: the first analysis provided an overall D score for CS material[48] , the second and third analyses provided separate D scores for English tokens and

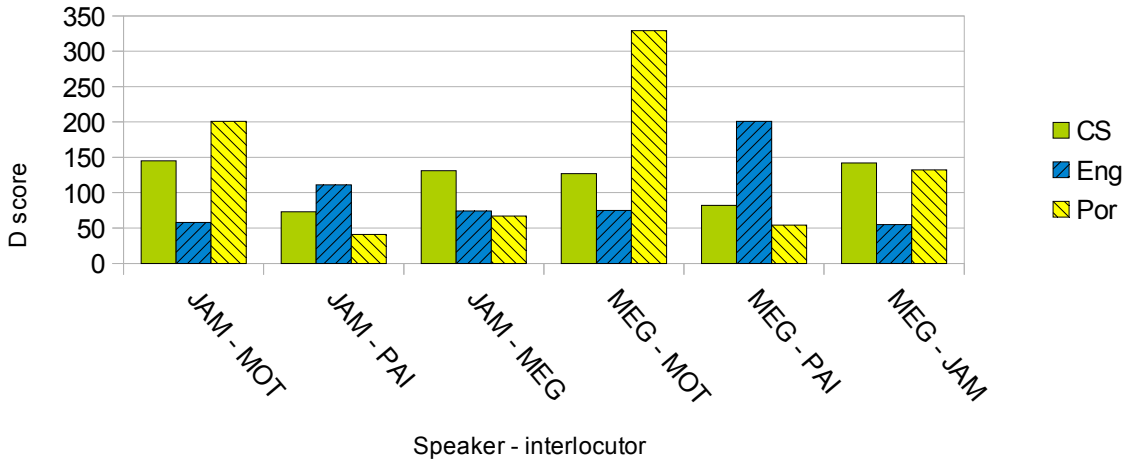
[48]kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]"

Portuguese tokens in the same material[49] . To aid interpretation of the results, the D scores pertaining to the siblings (as speakers) will be presented first, followed by those of their mother.

4.2.2.1 VOCD results of the siblings' code-switching with their parents and each other

The results of the batch of three VOCD analyses performed on six of the speaker/interlocutor combinations are shown in the chart below.

Figure 11. D scores for CS tokens and English and Portuguese tokens in CS utterances for JAM and MEG per addressee.



As was mentioned earlier, the reason for carrying out these VOCD analyses was to see if relationships could be found between the D scores and the specific roles of English and Portuguese in the code-switched discourse of the bilingual informants. If we examine the D scores of the siblings it is indeed possible to find such relationships but only because the variable of addressee has been isolated. If we compare the D scores that resulted from JAM and MEG's use of code-switching when addressing their mother (MOT) we find similarities. As indicated by the overall (non-addressee specific) CS D scores shown in Fig. 10, it is of little surprise that in Fig. 11 we find that both JAM and MEG have relatively high D scores for the CS speech addressed to their mother: 145 and 127 respectively. As discussed earlier,

[49] kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@pt>" and kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]"-s"<@en>"

these high scores are indicative of the lexical diversity found in these types of utterances. Of more interest here, however, are the separate D scores we find for English and Portuguese. For the English material in the CS utterances addressed to their English mother JAM has a D score of 58 and MEG has a score of 75. For the Portuguese material in the same CS utterances the D scores leap to 201 for JAM and 329 for MEG! This contrast in scores between languages is significant and reflects an asymmetry, in terms of lexical diversity, in the roles English and Portuguese play in the CS utterances.

Such asymmetry can also be identified from the results of the D scores of the siblings when the addressee is their Brazilian father (PAI). This time the D scores for English are higher than they are for Portuguese: 111 as opposed to 41 for JAM and 201 as opposed to 54 for MEG. When code-switching with their father, therefore, it is English which contributes a much wider range of lexical items to the CS utterances than Portuguese. And, as seen above, when code-switching with their mother the opposite is true: Portuguese provides far more lexical diversity than English. How does this relate to the Matrix/Embedded Language principle?

A hypothesis correlating D scores with the ML/EL asymmetry was proposed in section 3.3.5. It was suggested that in classic code-switching the typical grammatical nature of the contribution of the Matrix Language would result in relatively low D scores whereas the Embedded Language, through contributing content words, would be characterized by higher D scores. If we interpret the D scores in the chart according to this hypothesis, we are then able to identify whether English or Portuguese is acting as the ML or the EL in each case. Thus, when JAM and MEG address their mother, English (with relatively low D values of 58 and 75) is clearly taking on the role of the Matrix Language while Portuguese (with high values of 201 and 329) can be said to be acting as the Embedded Language. The roles are reversed when PAI is the addressee: Portuguese (with D scores of 41 for JAM and 54 for MEG) is the Matrix Language while English (with scores of 101 and 201 respectively) takes on the role of the Embedded Language. This interpretation supports the patterns found in the FREQ results (in 4.1.4) where the proportions of English and Portuguese tokens addressed to MOT and PAI by both siblings were found to reflect this differential use of the two languages: a higher token count equated with the Matrix Language whilst a relatively low token count correlated with the Embedded Language.

From the above discussion it does indeed appear that D scores could be a useful measure of differential language use in bilinguals, especially when examining their code-switched discourse. Having found such relationships between D scores and the ML/EL asymmetry it is now appropriate to examine the remaining results shown in Fig. 11, that is, the D scores of the code-switching which occurred in the interactions between the siblings. If we first look at the three D scores in the column 'JAM – MEG' (i.e. where JAM is addressing his sister) we find the following values: 131 for overall CS material, 74 for only English material and 67 for only Portuguese material. What we see here is a lack of disparity between the separate D scores for the two languages. According to my hypothesis this means that neither language can be deemed to be acting as the Matrix or Embedded Language: there appears to be more symmetry in their participation in code-switched utterances. This interpretation further supports the results provided by the FREQ analyses for the same speaker/interlocutor combination (see section 4.1.4). If we recall, the percentage of English to Portuguese words addressed by JAM to MEG was established as 43% to 57%, suggesting a more equal participation of both languages in the CS utterances. With very similar FREQ results for MEG (when addressing JAM) in terms of percentages (47% to 53%) we might expect the two separate D scores for English and Portuguese to also reflect this symmetry in language usage. However, if we look at the last column in the chart in Fig. 11 we find D scores which actually appear to indicate the sort of lexical asymmetry that was found in the CS discourse of both siblings when addressing their parents: English scores a relatively low 55, indicating its role as the Matrix Language, while Portuguese scores a high 132, reflecting a role more akin to the Embedded Language. As these VOCD results appear to be at odds with the findings discussed so far, I decided to examine the VOCD output for MEG in more detail, in search of a possible explanation.

In the output for any VOCD analysis, we are presented with not only the D scores but also the utterances which were used to calculate them. These are seen by simply scrolling up from the bottom of the output. In the case of the utterances addressed by MEG to JAM, what was found in the English only material was the repetitive use of the word 'the'. A simple FREQ analysis of the same utterances[50] revealed that this word occurred 14 times in a total token count of 91 (accounting for 15% of the tokens). On further examination of the utterances selected by VOCD it

[50]kwal @ +t%add +t*MEG +s"JAM" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5

could be seen that such frequent use of 'the' was actually due to the nature of some of MEG's interaction with JAM: she was reading a story to JAM in which 'the' appeared frequently, even more so due to her retracing of certain phrases. The repetitive use of one single word would certainly result in a lower D score but could this alone account for the D score of 55? In order to investigate this possibility, the same VOCD analysis as before was performed, but this time by adding the string **-s"the"**, all tokens of this word were removed from the input[51]. The resulting D score was certainly more in line with expectations, increasing to 106 (almost double the original D score). However, by removing tokens from MEG's input this makes her results less comparable to JAM's. To counteract this problem, exactly the same analysis (removing all tokens of 'the') was repeated on JAM's CS utterances addressed to MEG. His D score for English did increase slightly (from 74 to 88) but not as significantly as MEG's. It could be said then that her use of 'the' could be held partly responsible for the skewed results and that the asymmetry suggested by her D scores in Fig. 11 is not an accurate representation of the lexical roles English and Portuguese actually play in MEG's CS discourse with her brother. This observation hints at the effect interaction type (in this case reading a story aloud) can have on D scores, and indeed on other types of quantitative measures. This will be examined in section 4.2.3.

Of course, despite the D score purporting to account for the effect of size of input (ie number of tokens), it would appear logical that with very low numbers of tokens, such as 91 in the case of MEG above, any skewing of results through the repetition of certain words is naturally unavoidable. Of course, this investigation of the effect of 'the' on D scores was only carried out because the results were unexpected: they were not in line with the hypothesis that had been formulated between D scores and the Matrix/Embedded Language asymmetry. Whether or not the original hypothesis will be supported by future research carried out on other sets of data is of course important but within the confines of this study it has served to instigate further investigation.

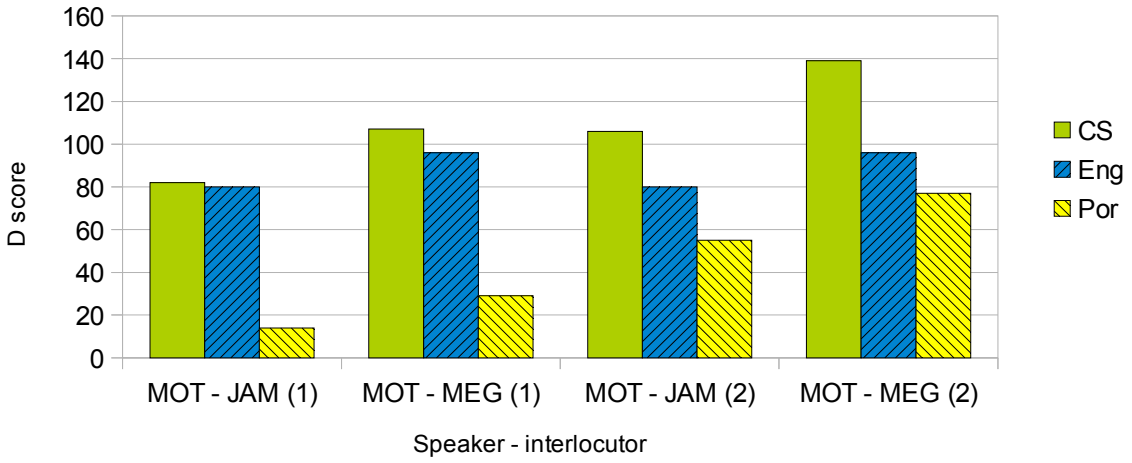
This need to delve futher into the results of the D scores also occurred when I examined those pertaining to MOT in her interactions with her children. This time, as will be seen below, an explanation for unexpected results did not lie with interaction type.

[51] kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@pt>" -s"the"

4.2.2.2 VOCD results of the mother's code-switching with her children

The D scores resulting from the VOCD analyses of the CS utterances addressed by MOT to her children were rather surprising. They are shown in the first two sets of columns (with (1) in brackets) in the chart below:

Figure 12. MOT's D scores for CS tokens, English CS tokens and Portuguese CS tokens in CS utterances: with 'olha' (1) and without 'olha' (2).



Before examining the separate D scores for English and Portuguese, if we look at the overall CS D scores for MOT's CS utterances (in (1)) we find relatively low scores: 82 when addressing JAM and 107 when addressing MEG. These are 'relatively low' when we consider that MOT's overall CS D score (non-addressee specific) reached 180 (see Fig. 10). However, it is the results for the English and Portuguese material that are more unexpected. According to my hypothesis, a low D score correlates with the Matrix Language while a high D score reflects the Embedded Language. If this is so, the D scores here are telling us that when MOT is interacting with both JAM and MEG, English is being used as the Embedded Language (with scores of 80 and 96) and Portuguese is taking on the role of the Matrix Language (with scores of 14 and 29). These results are surprising given that the FREQ results for MOT had established exactly the reverse! Further investigation was clearly necessary.

When examining the input used by VOCD to calculate these separate D scores for each language, nothing untoward was found in the English data. However, in the Portuguese data one word was found to appear extremely frequently and this word was 'olha', (in most cases reduced to 'o(lha)') which means 'look'. A quick FREQ

analysis[52] revealed that when MOT addressed JAM the repetitions of 'olha' (49) actually accounted for 37% of the total number of Portuguese CS tokens (132) and when addressing MEG the percentage was 24%, with 18 occurrences out of a total Portuguese CS token count of 75. To check whether such frequency of 'olha' could be found in the other bilingual speakers, the same FREQ analysis as above was performed on all bilingual speaker/addressee combinations (by simple speaker/addressee substitutions in the command line), giving rise to 10 additional analyses. The highest number of occurances found from these analyses was in JAM's code-switched utterances addressed to MEG where it occurred 5 times out of a total count of 164 for Portuguese CS tokens, that is, only 3%. Such FREQ analyses appear to suggest that MOT's use of 'olha' is idiosyncratic and not shared by her husband or children. Later qualitative analyses (see 7.1) will shed light on this particular usage.

For the present discussion what needs to be pointed out is the effect that such repetition of 'olha' has had on MOT's D scores for Portuguese, and subsequently on her overall CS D scores. In order to illustrate this I repeated two VOCD analyses on MOT's CS utterances addressed to her children (2 x 2 addressees), this time asking the programme to remove all occurrences of 'olha' from the input[53]. The results can be seen in the second two sets of columns in the chart above (marked as (2) in Fig. 12). The D scores for Portuguese appear to have risen quite significantly: from 14 to 55 when addressing JAM, and from 29 to 77 when addressing MEG. This has also caused MOT's overall CS D scores to increase: from 82 to 106 (to JAM) and from 107 to 139 (to MEG). This evidence strongly suggests that MOT's repetitive use of 'olha' has indeed had a significant impact on her D scores in Portuguese and subsequently on her overall CS D score and could partially explain why her results appear to contradict the hypothesis. However, it was not possible to perform a significance test in this case due to lack of data: attempts to output further D scores by breaking down MOT's data (for example a Portuguese D score per 10 files) led to error messages stating there were not enough tokens for VOCD to perform its analyses.

[52] kwal @ +t%add +t*MOT +s"JAM" +u +d | freq +o +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@en>"

[53] kwal @ +t%add +t*MOT +s"JAM" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"olha" and kwal @ +t %add +t*MOT +s"JAM" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@en>" -s"olha"

It is important to point out that even after having removed 'olha' from the input, the English D scores are still higher than the Portuguese D scores, albeit much less disparate. This still suggests that the English contribution to CS utterances is more lexically diverse than the Portuguese contribution, that is, that the former is playing a role more akin to the Embedded Language. Despite being at odds with what was established by the FREQ results it would be premature to throw out the hypothesis made between D scores and the Matrix/Embedded Language asymmetry. What has become clear through the investigation shown here is that quantitative results do not provide the whole picture and need to be interpreted with caution. The patterns revealed by quantitative analyses require further qualitative analysis if they are to be reliably interpreted.

Through the discussion of the VOCD results obtained from an analysis of the CS utterances of three of the bilingual informants, I hope to have shown that lexical diversity measures need not be restricted to the investigation of monolingual discourse. Indeed the analysis of the D scores discussed in this section has resulted in a novel hypothesis which correlates low and high D scores with the Matrix and Embedded Languages of code-switched discourse. While the results of the siblings (when addressing their parents) provide clear evidence for this hypothesis, those pertaining to their mother appear to contradict it. It is by investigating such contradictory evidence that we begin to learn more about the differences, and similarities, of the code-switching practice of each bilingual speaker.

The effect of idiosyncratic and contextual variables, such as interaction type, on the lexical diversity of CS discourse has been hinted at above, indicating that through analysing D scores we are potentially able to investigate such variables. Before leaving the discussion of the VOCD analyses performed on the LOBILL Corpus, I will consider two variables which I am able to investigate due to the diverse interactional and longitudinal nature of the Corpus, those of interaction type and time period.

4.2.3 VOCD analyses and results of the siblings' code-switching in different interaction types

It is of interest to this study that we examine the effect that interaction type can have on a speaker's use of code-switching and on the roles each participating language has to play in CS utterances. For example, when JAM is engaged in conversation at

the dinner table, do his CS utterances show a clear Matrix/Embedded asymmetry (revealed by his D scores)? And what about when playing board games: is a similar pattern found? Is his code-switched discourse more lexically diverse when chatting over the phone or when he is involved in face-to-face conversations? By using VOCD we are able to compare the lexical diversity found in the different types of interactions, something that could not be done (easily) via manual analysis of the transcripts.

It is important to mention that for the VOCD analyses reported on in this section, the input was selected by grouping the files according to interaction type. As each filename contains a two-letter code classifying its interaction type (see Table 7 below), this meant simply selecting the appropriate files from the drop down menu after clicking on 'File in' in the CLAN commands window. It was only necessary to perform this pre-selection of files once for each batch of VOCD analyses, the latter remaining exactly the same for each group of files.

The table below shows the seven types of interaction found in the LOBILL Corpus along with their classification code, number of files, overall token count, total CS token count and the percentage that this represents of the overall total. These token counts can be found at the end of the output provided by two general VOCD analyses of each interaction type[54].

Table 7. Token counts (overall and CS) per interaction type

Interaction type	Code	No. files	Overall no. tokens	No. CS tokens (% of overall token count)
Meal Times	MT	27	35455	2583 (7%)
Chatting	CH	20	26359	1865 (8%)
Playing Games	PG	15	26356	807 (3%)
Telephone Interactions	TI	25	17026	2414 (14%)
Literacy Activities	LA	11	14395	319 (2%)
Free Play	FP	14	12066	991 (8%)
Interviews	IN	6	5102	394 (8%)

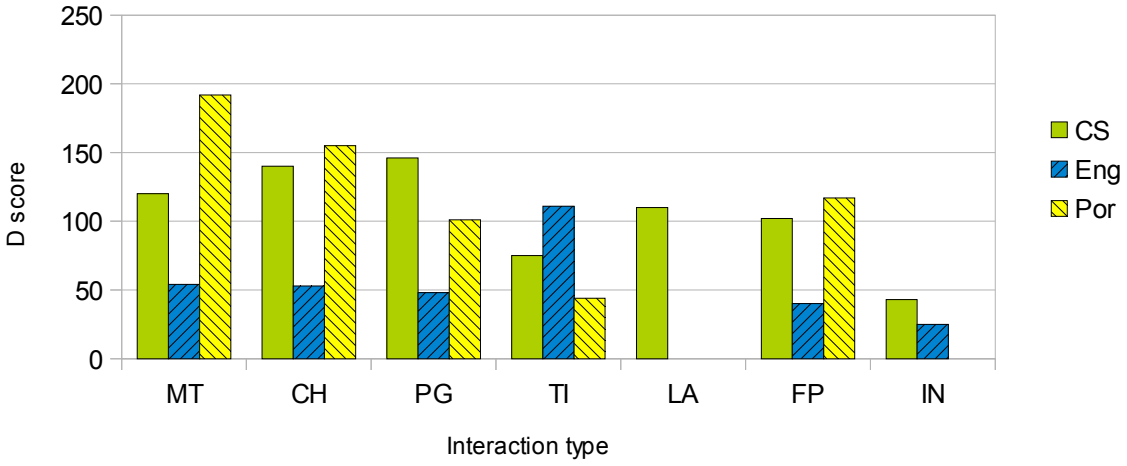
The interaction types have been ordered in terms of overall number of tokens and this is the order used in the charts presented below. Created for reference purposes,

[54]vocd @ +r5 -s"@nonwords.cut" +u and vocd @ +r5 -s"@nonwords.cut" +u +s"[+ *]"

there will be no discussion of the data in the table at this point except to remind the reader that despite the difference in numbers of tokens between the groups of files, VOCD is able to take these differences into account when calculating D scores.

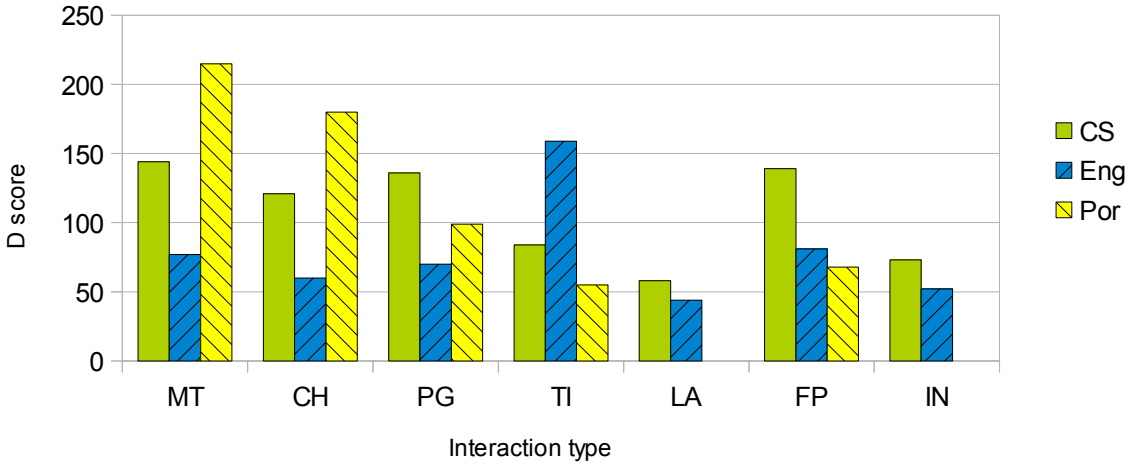
As for previous VOCD analyses, KWAL was used to select the speaker's CS utterances which were then analysed by VOCD to give an overall CS D score[55], a score for the English contribution to the utterances[56] and finally a D score for the participation of Portuguese[57]. Although these three analyses were performed on JAM, MEG *and* MOT's utterances, VOCD was only able to ouput sets of D scores for JAM and MEG. Low numbers of CS tokens for MOT meant that for almost all the seven interaction types there were D scores lacking for either Portuguese or English, or both. Thus, for comparative purposes, it only makes sense to present the D scores for JAM and MEG, shown below in Figs. 13 and 14.

Figure 13. JAM's D scores per interaction type for CS and English and Portuguese material within CS utterances



[55]kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]"
[56]kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@pt>"
[57]kwal @ +t*JAM +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@en>"

Figure 14. MEG's D scores per interaction type for CS and English and Portuguese material within CS utterances.



First of all, if we compare the two charts, what we find are very similar patterns. For all but two of the interaction types (the LA and FP groups) the D scores for both JAM and MEG indicate relatively the same variation in lexical diversity of English and Portuguese across the groups. This tells us that the siblings appear to be using both languages in similar ways in terms of the ML/EL asymmetry (or lack of).

For the TI (Telephone Interaction) group the high D scores for English and low scores for Portuguese indicate that in these interactions both JAM and MEG are using Portuguese as the Matrix Language and English as the Embedded Language. In the discussion of the FREQ results in Fig. 7, this particular combination of language roles was found to have occurred only in the interactions between the siblings and their Brazilian father (PAI). It seems logical to assume, therefore, that for the majority of the telephone interactions the siblings must be speaking to their Brazilian father and not their English mother. A quick FREQ analysis which outputs the number of times (in terms of utterances) each addresse is addressed in any given number of files[58] confirms that PAI was the main interlocutor of the telephone interactions (addressed 1497 times out of a total of 4784) while MOT was only spoken to on 321 occasions (see Table 8 below).

Table 8. Number of utterances (and % of overall total) addressed to the four bilingual informants per interaction type.

	MT	CH	PG	TI	LA	FP	IN

[58] freq @ +t%add -t* +u +o

MOT	3368 (37%)	2691 (38%)	3099 (34%)	321 (7%)	1896 (59%)	1298 (40%)	274 (40%)
MEG	2104 (23%)	1255 (18%)	2775 (30%)	859 (18%)	721 (22%)	727 (22%)	218 (32%)
JAM	2290 (25%)	2449 (34%)	2453 (27%)	820 (17%)	605 (19%)	1103 (34%)	190 (28%)
PAI	610 (6%)	88 (1%)	126 (1%)	1491 (31%)	3 (0.09%)	1 (0.03%)	2 (0.3%)
Overall total*	9175	7144	9170	4784	3233	3280	684

*these totals include utterances addressed to other speakers not included in the table

Apart from the TI group, judging by the number of times he is addressed in the other types of interactions, PAI's overall participation is clearly very low. This would help explain why the siblings' D scores for each language in the TI group do not follow those found in the other interaction groups: they reflect the fact that PAI's presence as main interlocutor has caused a language role reversal where Portuguese is playing a dominant role and English is contributing as an Embedded Language. If we look at the numbers in the Table for MOT and put aside the TI interactions (where only 7% of the utterances are addressed to her), we find consistently high proportions ranging from 34% (PG) to 59% (LA). That is, in each interaction type, MOT appears to be the main interlocutor, being addressed over a third of the time. This consistency in terms of being the most addressed speaker in all but the telephone interactions might go some way to explaining why the D scores for English for both siblings appear to vary relatively little across the groups. As both speaker and interlocutor, MOT appears to represent a constant in terms of language use (her FREQ results showed how consistently little Portuguese she used with her children) and this could be contributing to the lexical stability of English as indicated in Figs. 13 and 14. If this is so, it might be that the variability in Portuguese D scores across interaction types is actually reflecting a lack of consistency in terms of Portuguese-speaking interlocutors. In other words it may be that the sporadic presence of monolingual interlocutors alongside the four main informants is the cause for such variation in the Portuguese D scores. If this is the case, then it could mean that the siblings' use of Portuguese in CS utterances is not determined by the interaction type itself but differs according to who is present in the interaction.

Returning to the data shown in Figs. 13 and 14, there appears to be only one interaction type, FP, where JAM's D scores are not comparable to MEG's[59]. For the interactions in this Free Play group JAM appears to use English (with a low D score of 40) as the Matrix Language and Portuguese (with a relatively high score of 117) as the Embedded Language. In MEG's case the D scores are less disparate: 81 for English and 68 for Portuguese, indicating more symmetrical language use in her CS utterances for this interaction type. Only a more qualitative analysis of these particular interactions will shed light on the reasons for these differences between the siblings' D scores.

There is one final observation to be made when comparing the two charts: MEG's D scores for English are consistently higher than JAM's. The smallest difference is for the CH group where MEG's D score of 60 is only 7 more than JAM's score of 53. However, for the remaining five groups the difference ranges from 22 to 48, the latter being for the TI group. It is likely that MEG's relatively higher D scores for English are age-related: she is two and a half years older than James and as a result we can expect her lexical diversity to be comparatively higher. If we consider that for all but one of the interaction groups (the TI group), English is acting as the Matrix Language, one could infer that MEG's higher D scores, when compared to JAM's, are specifically reflecting her more diverse use of grammatical morphemes.

Such an interpretation would explain the consistency with which MEG's D scores for English are higher than JAM's. It would also go some way to explaining why the same consistency is not found when we compare the siblings' D scores for Portuguese: for two interaction types (PG and FP) JAM's D scores are higher than MEG's. Due to the nature of the typical contribution of the EL to CS utterances (mainly content morphemes) any grammatical diversity in Portuguese between the siblings would naturally be less in evidence (if evident at all). Any differences in D scores for the EL (Portuguese in this case), would therefore reflect diversity in terms of content morphemes, the use of which may be affected by contextal factors and not necessarily be related to linguistic development. The fact that JAM's D scores are higher than MEG's in two interaction types and the fact that there is such variablity in the Portuguese D scores across all the interaction types reinforces the notion that it

[59] The lack of D scores for the LA group (for JAM, English and Portuguese, and for MEG, Portuguese) means no comparison is possible here.

is necessary to take into account extra-linguistic factors when searching for explanations for CS patterns.

Isolating the variable of interaction type and establishing links between the code-switched discourse of the siblings occurring in each of the seven interaction types and lexical diversity is clearly a complex, and perhaps impossible, task. However, as seen above, in the process of examining the siblings' D scores per interaction type, significant differences in the roles of both languages in their CS utterances have been brought to the fore. The consistently lower values for English suggest two things: firstly, that across the interaction types (excluding TI) English is maintaining a stable participatory role in CS utterances; and secondly that this role, with its relatively low lexical diversity values, is most likely to be that of the Matrix Language. In contrast to the narrow range of D scores found for English, the Portuguese D scores revealed a wide variation in lexical diversity across the different groups implying that Portuguese is contributing in different ways to CS utterances. It may be that the structured nature of game playing, where repetition is common, is the cause of relatively low D scores for Portuguese (and English). It follows that we might expect to find higher D scores for Portuguese in less structured interaction types such as meal times, chatting and free play activities. In terms of the relationship between D scores and the roles of the ML/EL, the results do suggest that the latter types of interaction are more conducive to the typical ML/EL patterns found in classic code-switching.

Further findings from the analysis of the data highlighted the need to take into account developmental differences when interpreting the D scores of the children, especially with regards to the Matrix Language: MEG's consistently higher D scores in English when compared to JAM's were interpreted as being the result of her use of a wider variety of grammatical morphemes, due to her higher (age-related) level of grammatical competence. However, variability between the siblings in their D scores for Portuguese indicated that as far as the Embedded Language was concerned, differences in linguistic development were not the only factors to be considered. Although the similarity in the patterning of the siblings' Portuguese D scores across six of the interaction types did seem to suggest a relationship between interaction type and the lexical diversity of the Embedded Language, it was pointed out that such patterns could equally be due to the interlocutor variable. Unfortunately, when I introduced this addressee variable into the previous analyses, in most cases there

were insufficient utterances for VOCD to output D scores separately for each language per interaction type.

Although it was not possible to combine the above two variables (interaction type and addressee) in the VOCD analyses, it did prove possible to combine the addressee variable with that of the variable of time. The discussion of these analyses will be the focus of the next section.

4.2.4 VOCD and a longitudinal analysis of the siblings' code-switching with their mother

The files in the LOBILL Corpus cover the period of time from August 2001 to December 2004. Although the recordings are not equally spread over this time span, leading to unequal numbers of tokens per month, as mentioned before, VOCD is able to take this into account and the D scores should still reveal any differences found in lexical diversity across the time period. Whereas in the previous section the files were divided into seven different groups according to interaction type, this time the files were divided up longitudinally. Various experimental VOCD analyses were carried out to see how the files could be grouped for maximum effect, that is, the minimum number of files per time period which would still allow VOCD to output D scores. As I wished to control for speaker and addressee variables, this meant that it was only the interactions between the siblings and their mother that provided sufficient code-switched material for VOCD to perform the analyses.

I decided to divide up the 119 files of the corpus into eighteen different groups (see Table 9 below), each one with a minimum of three files and a maximum of eight. An attempt was made to ensure that there were no major disparities in the total number of tokens for JAM and MEG across the longitudinal groups (these totals will be considered in the discussion of the results). This meant that each group of files does not span exactly the same time period, especially as in some months recordings were carried out more frequently (for example when on holiday in England). It was also considered appropriate to align the division of the files with changes in location: groups 8, 9 and 10 cover the period when the siblings were in England on holiday (June and July 2003); and groups 16, 17 and 18 include the recordings that were carried out after their move to England (in June 2004). Although this section is concerned with charting D scores longitudinally, as will be seen, it is not the effect of time itself that is the focus of the investigation but rather how the D

scores are affected by time-related events and contextual changes occurring over the three years of the siblings' bilingual language journey.

In previous discussions it has proven useful to triangulate VOCD results with FREQ results, especially when looking for evidence which might support my hypothesis which proposes a relationship between high frequency word counts/low D scores and the Matrix Language and between low frequency word counts/high D scores and the Embedded Language. Although evidence contrary to my hypothesis has emerged, this has served to provoke further investigation which will ultimately allow for more focussed qualitative analysis in later chapters. It is for this reason, that in this section the results for both D scores and related CS token counts will be presented together, allowing for direct comparison. An added advantage of being able to access information about the token count for each language is when VOCD is unable to provide D scores: the separate token counts still indicate which language is participating more in the CS utterances.

Before detailing the analyses performed, I will first provide an overview of the longitudinal group divisions for reference purposes. The table below shows the group number, file numbers, time period, speaker token totals (*all* tokens addressed to MOT by JAM and MEG) and interlocutors for each group. In order to output the total number of tokens addressed by JAM and MEG to MOT per group (fourth column), the files were pre-selected from the drop down menu and then KWAL was used to specify the input which was then sent to VOCD[60]. To output the interlocutors featuring in each group a simple FREQ analysis was used[61]. The output provides the number of times each speaker code can be found on the dependent tier **%add**, i.e the number of times each participant is addressed. Although these totals are not displayed in Table 9, they will be used in the discussion of the results in 4.2.4.3. The interlocutors appear in column 5, bilingual interloctors in bold font, monolingual English interlocutors in italics and monolingual Brazilian interlocutors in normal font.

Table 9. Longitudinal division of LOBILL Corpus for VOCD analyses.

Group	Files	Time period	Token total (addressed to MOT)	Interlocutors (in order of most addressed)
1	001-007	2001 AUG-NOV	JAM: 467 MEG: 1341	**MOT**, **MEG**, **JAM**, *BEC*, *GRA*, SAR, **PAI**

[60]kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut"
[61]freq @ +t%add -t* +u +o

2	008-016	2002 JUN/JUL	JAM: 958 MEG: 1515	**MOT**, **MEG**, **JAM**, **PAI**
3	017-023	JUL/AUG	JAM: 1431 MEG: 2078	**MOT**, **JAM**, **MEG**, **PAI**
4	024-030	OCT-DEC	JAM: 1482 MEG: 2295	**MOT**, **JAM**, **MEG**, *GRA*, **PAI**
5	031-037	2003 JAN-MAR	JAM: 1505 MEG: 1000	**MOT**, **JAM**, **MEG**, SAR, *GRA*, **PAI**, AVO, JUL
6	038-044	APR	JAM: 1004 MEG: 2416	**MOT**, **MEG**, **JAM**, **PAI**
7	045-052	APR/MAY	JAM: 1035 MEG: 3115	**MOT**, **MEG**, **JAM**, SAR, **PAI**, JUL, AVO
8	053-060	JUN/JUL	JAM: 1613 MEG: 1717	**MOT**, **MEG**, **JAM**, *BEC*, *JAK*, *MAX*, *GRA*, **PAI**, VIN
9	061-068	JUL	JAM: 1613 MEG: 3177	**MOT**, **JAM**, **MEG**, **PAI**, *JAK*, *MAX*, *BEC*, *GRA*
10	069-075	AUG	JAM:438 MEG: 1545	**MOT**, *WIL*, **PAI**, **MEG**, **JAM**, *GRA*
11	076-080	AUG/SEP	JAM: 2164 MEG: 1559	**MOT**, **JAM**, **MEG**, **PAI**, ARL
12	081-085	OCT	JAM: 1323 MEG: 1614	**MOT**, **JAM**, **MEG**, **PAI**
13	086-091	NOV	JAM: 1322 MEG: 1081	**MOT**, **JAM**, **MEG**, **PAI**
14	092-094	DEC	JAM: 427 MEG: 156	**JAM**, **MOT**, **MEG**, *GRD*, *GRA*
15	095-099	2004 MAR-JUN	JAM: 764 MEG: 1395	**MOT**, **MEG**, **JAM**, **PAI**, *GRA*
16	100-106	JUN/JUL	JAM: 369 MEG: 563	**MEG**, **JAM**, **PAI**, **MOT**, AVO, VOV, *GRA*
17	107-111	AUG	JAM: 802 MEG: 972	**MOT**, **JAM**, **MEG**, **PAI**, AVO, VOV
18	112-119	OCT-DEC	JAM: 837 MEG: 1321	**JAM**, **MEG**, **MOT**, AVO, VIN, SAR, VOV, JAN, ROS, **PAI**, DAN

Throughout the following discussion, reference will be made to the information in this table as we search for possible explanations for the patterns found in the results.

The results which are shown in the four charts (see below) were achieved by performing two VOCD analyses repeatedly on JAM and MEG's CS utterances for each of the eighteen groups of files. One of the analyses outputted the D scores and token count for the English material found in the CS utterances addressed to MOT[62] while the other provided the same type of output for the Portuguese material[63]. For both JAM and MEG, therefore, we have two charts each, the first ones (Figs. 15 and 16) depicting the token counts and the second two (Figs. 17 and 18) showing their D

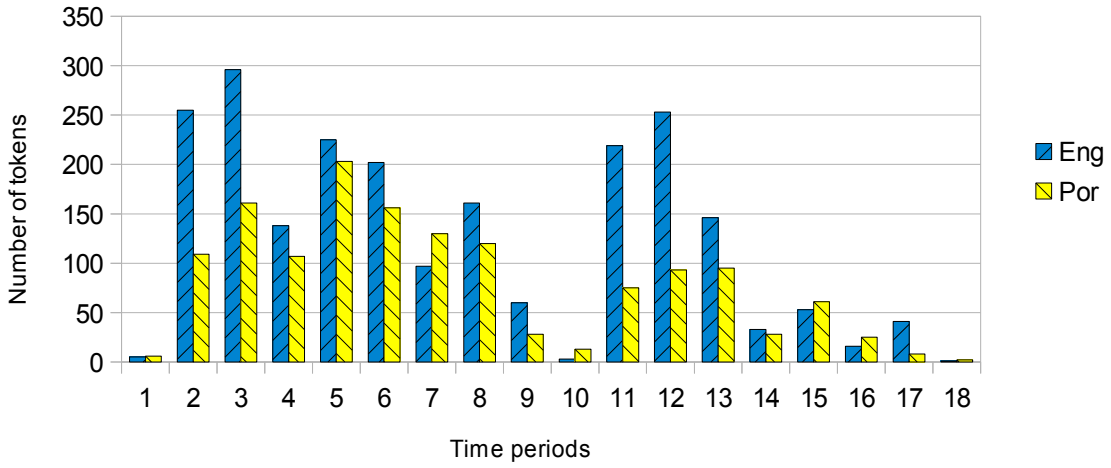
[62]kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@pt>"
[63]kwal @ +t%add +t*JAM +s"MOT" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@en>"

scores. After making separate observations on the data in the two types of charts, relationships between token counts and D scores will then be examined and discussed in the light of contextual information such as location and interlocutor presence which were seen to differ over the time periods. Comparisons will also be made between the siblings to shed further light on the similarities and differences which have emerged through previous findings.

4.2.4.1 Token counts for the siblings across the time periods

In this section I examine the separate token counts for English and Portuguese in JAM and MEG's CS utterances addressed to MOT. Looking first at the results for JAM in the chart below, one can see wide variation in the numbers of tokens across the time periods.

Figure 15. Number of tokens per time period for the English and Portuguese material in CS utterances addressed by JAM to MOT.



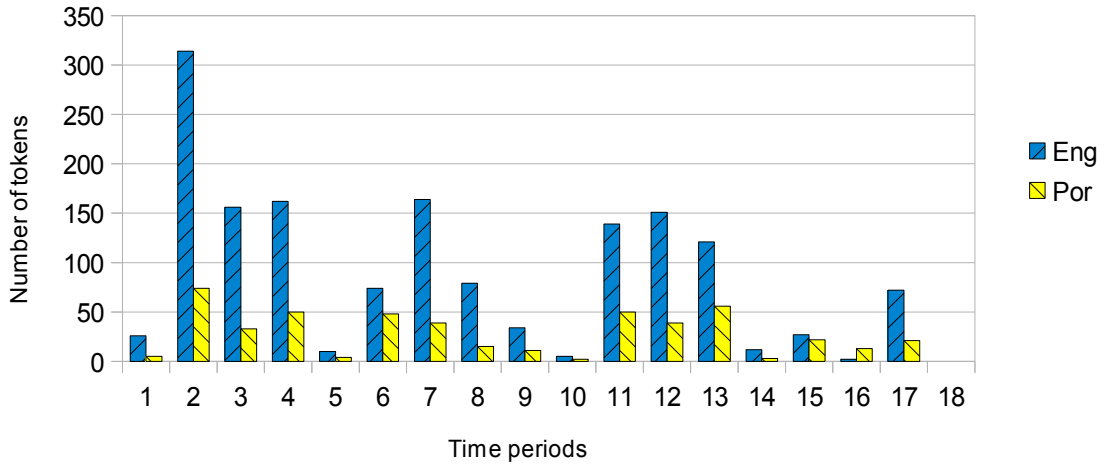
Although we would not expect exactly the same number of CS tokens across the groups (due to differences in overall token counts), if we compare some of the groups it becomes evident that even taking total tokens into account, there appear to be significant differences in the amount of code-switching that JAM is engaging in with his mother. For example, for groups 1 and 10 we find very low numbers of English and Portuguese CS tokens (5/6 for 1 and 3/13 for 10). This is not surprising considering JAM's total token account for these two groups: 467 for 1 and 438 for 10 (see Table 9). However, if we then compare groups 3 and 9 we find a disparity which cannot be put down to differences in total token size. Despite a higher token count of

1613 for group 9 when compared with the 1431 tokens for group 3, it is in the latter group that we find much more code-switching taking place, reflected in the higher number of CS tokens: 457 for group 3 as opposed to 88 for group 9. This means that over the time periods JAM is not showing consistency in terms of the amount of code-switching he uses when addressing his mother. What are the factors affecting his usage? This will be returned to later on in the discussion.

When it comes to consistency in terms of the roles each language plays in JAM's CS utterances addressed to MOT, the results do appear to reveal a tendancy which had become evident in earlier FREQ and VOCD results: that English plays a more dominant role than Portuguese. In 12 out of the eighteen groups, the token count for English is higher than that for Portuguese. Of these 12 groups there are six (2, 3, 9, 11, 12 and 17) where the wider relative difference between the numbers of English and Portuguese tokens reflects the typical asymmetry found in 'classic' code-switching, English taking on the role of Matrix Language and Portuguese the role of Embedded Language. For the other six groups (4, 5, 6, 8, 13 and 14) the difference is less marked, although English still plays a more dominant role. But what of the remaining six groups where the token count for Portuguese is higher than that for English (groups 1, 7, 10, 15, 16 and 18)? What could be causing Portuguese to take on a more active role in these groups? If we recall, JAM is still addressing the same interlocutor, his mother, with whom, the evidence suggests thus far, he favours English when code-switching. There must be other factors affecting his increase in use of Portuguese in these time periods.

In terms of any changes in language dominance over time, on first examination, there appears to be no particular pattern: the CS tokens for each language do not, for example, show a progression over time from classic CS to utterances in which both languages play an equal role. What is perceptible, however, are two stretches of time where JAM's total CS count seems to dip quite dramatically: in time periods 9 and 10 and between 14 and 18. Before I look at the time-related contextual factors which could have caused these dips, we will first examine MEG's token counts and compare them to JAM's.

Figure 16. Number of tokens per time periods for the English and Portuguese material in CS utterances addressed by MEG to MOT



What is immediately apparent in MEG's results is the consistency with which English plays a more dominant role in her code-switching with her mother across the time periods. There is only one time period where Portuguese participates more than English and that is in group 16 for which the token count is very low: 2 tokens for English and 13 for Portuguese.

Classic code-switching seems to be the style favoured by MEG, as can be seen from the high disparity between English and Portuguese tokens in almost all time periods. There are only 4 time periods (5, 6, 10 and 15) where the difference in token counts is less disparate. If my hypothesis is correct, Portuguese is clearly taking on the role of Embedded Language in most CS interactions between MEG and MOT. It is evident from the chart that MEG's total tokens for both English and Portuguese are consistently lower than JAM's over the time periods. However, we are still able to find the same peaks and troughs in the charts which indicate that contextual features are affecting JAM's and MEG's code-switching in a similar fashion. The only major difference is for time period 5 where JAM's CS token counts for English and Portuguese are 225 and 203 as opposed to 10 and 4 in MEG's case. A possible explanation for this difference will be sought in the contextual information we have for each time period. Firstly, however, the data from these two charts (Figs. 15 and 16) will be triangulated with the results of the D scores shown in the next section (Figs. 17 and 18). It is expected that the D scores for both JAM and MEG will provide supporting evidence for the observations made above regarding the differing

roles of English and Portuguese in the CS utterances addressed to MOT by the siblings.

4.2.4.2 D scores for the siblings across the time periods

If we recall, the hypothesis is that high D scores reflect the lexical diversity typically found in the Embedded Language while relatively low scores would characterize the contribution of the Matrix Language. If this is so, despite the different scales used in the y axis of the charts (number of tokens and D scores), we should find that to some extent, the patterns in the D scores charts visually reflect those we see in the token charts, the only difference being a reversal in terms of language. For example, for periods 2 and 3 (with *high* English/*low* Portuguese token counts) we might expect to find *low* D scores for English and *high* D scores for Portuguese. Where the token counts for both languages were less disparate (such as in 5 and 6) we might expect to find more equal D scores. Clearly, as was seen in the section on interaction types, the D scores can be affected by the nature of the interaction itself whereas token counts are immune to this effect. However, for the moment we will see if, in general terms, the predictions mentioned above are reflected in the two charts discussed below (Figs. 17 and 18).

Figure 17. D scores per time periods for the English and Portuguese material in CS utterances addressed by JAM to MOT.

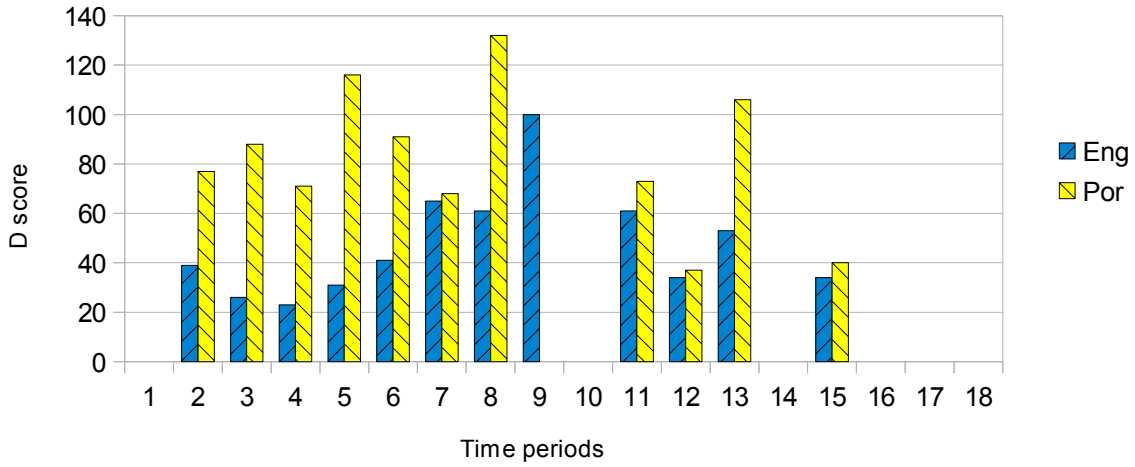
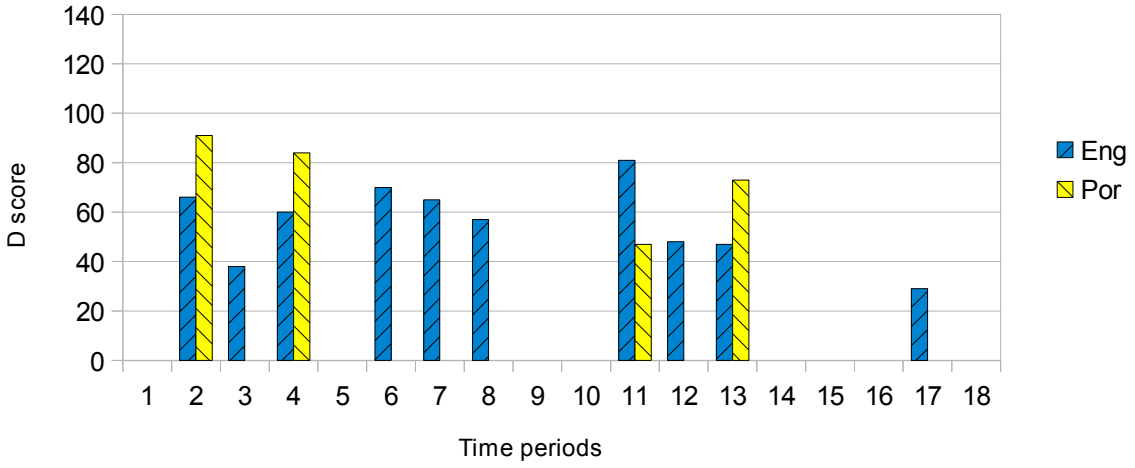


Figure 18. D scores per time periods for the English and Portuguese material in CS utterances addressed by MEG to MOT.



First of all, what one notices is the absence of D scores for some time periods. For both JAM and MEG there are no D scores at all for periods 1, 10, 14, 16, and 18. And while JAM is also missing a D score for 17, MEG is missing D scores for 5, 9 and 15.

If VOCD was unable to provide D scores for these periods this means that in each case there were less than 50 tokens for each language for that period. That is, very little code-switching took place. While for JAM there is only one other occasion where a D score is missing (for Portuguese in group 9), for MEG there are six other D scores missing, all for Portuguese, in 3, 6, 7, 8, 12 and 17. This lack of D scores

for MEG tells us that she uses Portuguese sparingly when code-switching with her MOT. These findings corroborate the low token counts for MEG for these periods (3, 6, 7, 8, 12 and 17) shown in Fig. 16 (all under 50 tokens). It is important to highlight this reliability in terms of the functioning of the VOCD programme as it means that one can reliably infer from future VOCD analyses that an output of *no* D score automatically means that the occurrence of CS (whether examining CS material as a whole or per language) amounts to less than 50 tokens - there would be no need for further investigation via FREQ to verify any absence of D scores. Putting this methodological consideration aside, let us consider the predictions mentioned above and examine whether, despite the missing D scores, relationships between D scores and the roles of the languages can be found.

When we look at JAM's D scores in Fig. 17, what we see is that for every time period, without exception, Portuguese scores more highly than English. For the majority of time periods (2, 3, 4, 5, 6, 8 and 13) the difference in scores between the two languages is more than 35. However, there are four groups (7, 11, 12 and 15) where the scores are less disparate. In general terms it appears that the pattern of D scores supports the prediction that we might expect higher scores for Portuguese, given the relatively low token counts for Portuguese shown in Fig. 15. These findings combined provide strong evidence for the asymmetrical roles the languages play in the majority of JAM's CS utterances addressed to MOT: English acts as the Matrix Language while Portuguese seems to conform to the role of Embedded Language. There *are* cases where the D scores appear to be slightly at odds with what the token counts for each language would predict: the slightly higher token counts for Portuguese in time periods 7 and 15 do not result in slightly lower D scores for Portuguese when compared to English. However, the token results of these two particular groups (and groups 1, 10, 16 and 18) had already been highlighted as showing a certain divergence from the 'normal' pattern of language participation in JAM's CS utterances. Clearly such groups need to be the focus of further investigation.

With regards to the D scores in MEG's chart (Fig. 18), there are only four time periods (2, 4, 11 and 13) where we have scores for both English and Portuguese. Three of these four (2, 4 and 13) follow the pattern found for JAM: relatively lower scores for English and higher for Portuguese. In MEG's case, however, there is less disparity between the separate English/Portuguese scores (66/71 for group 2; 60/84

for group 4 and 47/73 for group 13), averaging a difference of only 25. For JAM this average difference is 46. If we look at all the D scores for English in MEG's chart, we find that for all but two of the time periods, MEG's scores are consistently higher than JAM's. This would bring her D scores for English closer to those for Portuguese, thus resulting in a smaller average difference. These findings support what was shown by the VOCD analyses carried out per interaction type (see Figs. 13 and 14): there as well, MEG's D scores for English were found to be consistently higher than JAM's.

What does this mean in terms of my hypothesis about the relationship between low D scores and the Matrix Language and between high D scores and the Embedded Language? It is becoming evident that this hypothesis is rather simplistic and does not take into account differences in lexical diversity which are the result of age-related linguistic development. In its original form the hypothesis would interpret MEG's relative lack of disparity between the D scores of both languages (in groups 2, 4 and 13) as indicating that her code-switching with the mother involved more equal participation of English and Portuguese. However, the token scores seem to indicate that the style of her code-switching is definitely classic, that is, there is clear asymmetry in her use of both languages in CS utterances. Evidently the D scores hypothesis needs to take into account the natural increase in lexical diversity which occurs as children develop linguistically over time. By incorporating this variable, the original hypothesis would provide a more accurate method for interpreting the D scores of code-switched utterances in terms the of Matrix and Embedded Language asymmetry. This issue will be discussed further in Chapter 8.

Returning once more to MEG's chart, there is one more time period which deserves special mention owing to its apparent anomaly. For time period 11, English scores significantly higher than Portuguese in terms of lexical diversity, 81 as opposed to 47. This implies that MEG is using Portuguese as the Matrix Language when code-switching with her mother. This is unexpected, especially as the token count of 50 for Portuguese and 139 for English (see Fig. 16) indicate exactly the opposite, that English is the Matrix Language! It is only through the triangulation of results that such an anomaly has come to the fore, demonstrating how important it is to combine different types of measures when investigating via quantitative methods.

In the discussion of the results shown in the four charts, comparisons have been made between the token counts and D scores for each sibling. Comparisons have also been made between the siblings in order to highlight any similarities and

differences in their use of English and Portuguese in the CS utterances addressed to their mother. For both JAM and MOT, certain time groups have been singled out as warranting further investigation in order to explain why these particular groups do not appear to follow suit in terms of expected token counts and D scores for each language. Such investigation will involve looking at the contextual information for each time period, some of which is available in the table which features at the beginning of this section (Table 9). Additional information about location and interaction type can be found in the File list (see Appendix A). It makes sense to present the discussion in chronological order: this way any differences over time will become more apparent.

4.2.4.3 Factors affecting the siblings' token counts and D scores across the 18 time periods

We saw from the token and D scores charts that both JAM and MEG engaged in very little code-switching with their mother in period 1. A possible explanation for this can be found in the interlocutor column in Table 9: two monolingual speakers, GRA and BEC were the fourth and fifth most addressed interloctuors in this time period. The siblings' English grandmother and aunt were on a visit to Brazil and feature heavily in the interactions. It appears that the presence of these two monolingual speakers may have caused the siblings to reduce their use of code-switching. This reduction may have been due to a conscious effort on the part of the siblings to accomodate their speech so that their relatives would not be excluded from the conversations. However, it might also have been just the natural result of the increased dominance of English in this time period. Qualitative analyses of the utterances will reveal more.

In contrast to period 1, the results of the VOCD analyses of periods 2 and 3 revealed substantial use of CS by both siblings when addressing their mother. The examination of the token counts and D scores showed how consistently English was being used as the Matrix Language and Portuguese as the Embedded Language. Such consistency could not have been due to interaction type as the files in these periods (008-016 and 017-023) include six out of seven of the interaction types. The only constant variable in these time periods are the interlocutors, the siblings' parents, MOT and PAI. If we compare the number of times each parent is addressed, we see that PAI actually plays a very small part in these interactions: in period 2 he is addressed only 54 times out of the total of 1223 and in period 3 this falls to only 5

times out of a higher total of 2747[64]. It is the mother who is addressed the most: 548 times for period 2 and 1025 times for period 3. With no monolingual interlocutors present the siblings would have no need to restrict their use of code-switching.

The pattern of code-switching found for period 4 is very similar to that of the two previous periods. This may seem surprising when we see from the table that GRA (the English grandmother) features in the interlocutor list. However, out of a total of 3585 times, GRA is only addressed 262 times (7.3%) and we learn that this is concentrated in only one of the interactions (file 030) when she pays a christmas visit to Brazil. In all of the other interactions (024-029) MOT represents a constant in terms of interlocutor and therefore the impact of GRA's presence may not be enough to be felt in terms of token counts or D scores.

It is when we come to time period 5 that we see a divergence in the amount of CS each of the siblings use with their mother. The token counts and (lack of) D scores had shown that MEG's CS was virtually non-existent, amounting to a mere 14 CS tokens. This contrasted quite dramatically with JAM whose total CS token count of 426 was seen to be relatively equally split between English (225 tokens) and Portuguese (203 tokens). This more equal participation, in terms of numbers of tokens, was not mirrored in the asymmetrical D scores where a high score of 116 for Portuguese contrasted with 31 for English. If we examine the interlocutor column in the table we find three monolingual Portuguese speakers (SAR, AVO and JUL) and one monolingual English speaker (GRA). Could it be that the presence of these Portuguese speakers has resulted in an increased participation of Portuguese in his CS utterances with his mother? From the Portuguese D scores it would seem that this increase is in the use of Embedded Language items, given the high lexical diversity. But what of the English grandmother's presence in two of the files (031 and 032)? Was her presence not strong enough to have had an influence on the amount of CS JAM used with his mother? It appears that despite the presence of these monolingual speakers, JAM does not reduce his use of CS this time (compare this to period 1 above). We would need to look at the utterances themselves to see what is happening in these interactions. By doing this it would also be possible to investigate whether MEG's noticeable lack of CS is due primarily to her accommodation of her interlocutors' monolingualism.

[64] These and the following totals which relate to the number of times an interlocutor is addressed were provided in the output of the frequency analysis shown in footnote 61.

In terms of interaction type and interlocutors, time period 6 is comparable to time periods 2 and 3. The amount and pattern of JAM and MEG's CS is also comparable, although MEG appears to code-switch slightly less, and JAM's use of Portuguese seems to have increased (when compared to his use of English) .

For period 7 we again find differences between the siblings' use of CS. While MEG maintains the classic asymmetrical type of CS with her mother, JAM's increased use of Portuguese is actually indicating that this language is starting to vy for the role of the Matrix Language. There are actually more Portuguese tokens than English and his D scores reflect approximately equal lexical diversity. The presence of three monolingual Portuguese speakers in two of the interactions may have contributed to this increased participation of Portuguese in JAM's CS with his mother. However, in the previous time period (6), where there were no monolingual interlocutors present, an increased participation of Portuguese had already been noted. Could it be that the dominance of Portuguese in his environment outside of the family home is beginning to become more influential in his bilingual language use? If this is the case, MEG does not seem to be as susceptible as JAM to this linguistic influence: her CS token count for Portuguese was less than 50 tokens for this period.

Due to a dramatic change in the linguistic environment occurring in period 8, it was not possible to track JAM's increased use of Portuguese to see whether this language would eventually become the Matrix Language in interactions with his mother. All the files in periods 8, 9 and 10 (Files 053-075) contain recordings made in England where the siblings and their mother were on holiday for two months. Evidently, immersion in a purely monolingual English environment will have an effect on a bilingual's language use but let us recall that we are still analysing the siblings' CS utterances with their *mother* whose language use has been consistent. Therefore, here we are looking primarily at the influence of the variables of location and the presence of other interlocutors.

For period 8, JAM appears to revert back to the classic asymmetical pattern of CS which was found in the earlier time periods, English reafirming its role as the Matrix Language. Despite the presence of monolingual English speakers (BEC, JAK, MAX and GRA) in the interactions for this time period, JAM still engages in a substantial amount of code-switching with his mother while MEG's code-switching seems to be more measured. As we progress into period 9 (a month after the

siblings' arrival in England) we see less use of CS by both JAM and MEG and by the time we reach time period 10 (seven weeks after having arrived), there are very few CS tokens indeed: 16 for JAM and 7 for MEG. In this last period the siblings' English uncle (WIL) features as the second most addressed interlocutor (see Table 9) and the siblings may have had to accommodate the language addressed to their mother due to his presence. However, it seems likely that after two months of English immersion, there was no need to call on Portuguese for lexical gap reasons: English was actively providing all the necessary linguistic means for communication, especially for those experiences particularly bound by the social and cultural environment in England. For example, activities such as going to the local library or travelling by train were not experiences that had been part of the siblings' life in Brazil. Thus Portuguese would never have been a natural choice to talk about such activities with their mother.

Time period 11 sees JAM and MEG back in Brazil and back to their normal routines. This applies to their linguistic routine as well and we see a return to the classic use of CS with their mother in time periods 11, 12 and 13. The increased participation of Portuguese which had been noted in JAM's CS before his trip to England seems to have been annuled by his English immersion. There is one difference that is worthy of note and that is MEG's D score for English in period 11 (which covers the month after her return). Unexpectedly, we find a score of 81 which is almost double that for Portuguese (47), implying that English, instead of Portuguese, may be responsible for the lexically rich contribution to CS utterances, normally typical of the Embedded Language. Although it is perfectly feasible that her overall lexical diversity in English must have been given a boost through her stay in England, it is not logical to accept that MEG is now using English as the Embedded Language and Portuguese as the Matrix Language. Indeed, the token counts support exactly the opposite and are representative of MEG's 'normal' CS practice with her mother. Closer qualitative analysis of the utterances for this time period will reveal what is causing this reversal of D scores.

There is not much to be said about period 14 as very little CS occurred (60 CS tokens for JAM and 15 for MEG). Although we could attribute this to the presence of two monolingual English speakers (GRA and GRD), it is more likely that the CS token counts are low due to the low number of overall tokens uttered by both siblings to their mother: 427 for JAM and 156 for MEG. If we also consider that JAM and

MEG were addressing their English grandmother and grandfather over the phone this means that the latter were not really 'present' in that sense. Any utterances addressed to MOT would not have been heard by the English interlocutors anyway so any linguistic accommodation by JAM and MEG would have been unnecessary.

For period 15 the CS token counts for both JAM and MEG are also very low (114 and 49 respectively). With no D scores for MEG, all that can be said for her use of CS with her mother is that it is minimal for this period. With regards to JAM, despite the low token count, Portuguese again appears to be gaining ground in the CS utterances with 61 tokens as opposed to 53 for English and with almost equal D scores, indicating more equal participation of both languages. Although GRA appears in the interlocutor list for this period, again it is as a telephone interlocutor so her 'presence' is unlikely to have affected JAM's CS use with MOT.

Period 16 marks a new phase in the siblings' lives with their permanent move to England. Very low CS token counts are found for both children, only slightly more than period 10 when CS had virtually disappeared from the utterances addressed to their mother, just before their return to Brazil. Period 16 should really be comparable to period 8 as these both mark the start of an abrupt change in linguistic environment. However, the gradual decrease in CS seen over periods 8 and 9 is not reflected in the data for period 16. It seems plausible to suggest that their two months of linguistic experience in England the previous year had served to prime them for when they returned 10 months later.

It is interesting to note that, despite the very low token counts, there are relatively more Portuguese tokens than English (25 versus 16 for JAM and 13 versus 2 for MEG). This may have something to do with the interaction types found for this period, four of which involved telephone interactions in Portuguese between the siblings and their father (PAI) and their grandfather (VOV). Although JAM and MEG were speaking to their father, as will be seen later in this chapter, both siblings, but especially JAM, would turn to MOT for assistance when having difficulty talking about a typically English concept. This may have resulted in more Portuguese tokens but only an examination of these interactions will shed light on this supposition. During period 17 the siblings' father (PAI) joins his family in England and adds another dimension to the linguistic dynamics of the home. It may be that his bilingual presence in the last two interactions of this time period results in the slight increase in CS tokens addressed to MOT that is found for both JAM and MEG, this time with

English firmly in the role of the Matrix Language and Portuguese as the Embedded Language. Interestingly it is for MEG that we have the higher token count, even enough to provide us with a D score of 29 for English, such a score reflecting the (lack of) lexical diversity of a Matrix Language.

Between periods 17 and 18 there was a month where no recordings were made. This month (September) may have revealed a gradual decrease in the amount of code-switching used by JAM and MEG when addressing their mother. Although this is inferred, it supports what the results for period 18 show - a complete lack of CS for both siblings. When looking at the interaction types, we discover that six out of eight are telephone interactions between JAM and MEG and various Brazilian relatives and friends. Perhaps the siblings did not actually interact much with their mother. However, the total tokens column in the table show that JAM addressed 837 tokens to his mother while MEG's total token count is higher, at 1321. Is this absence of code-switching in period 18 representative of what their continuing CS practice with MOT would have looked like, i.e. non-existent? And what about when addressing PAI? The interactions in this period need to be examined very closely in order to verify whether any other code-switching takes place, either between the siblings and their father or between themselves.

By carrying out this quantitative longitudinal analysis of the siblings' CS utterances addressed to their mother, one of the aims has been to demonstrate how quantitative measures can be used to investigate the roles of the participating languages over time. The token counts and D scores outputted by VOCD have revealed differences and similarities between the siblings in their use of CS with MOT. The longitudinal division of the LOBILL Corpus has made it possible to search for relationships between these results and contextual variables such as interlocutor presence and location which are ultimately linked to particular time periods. The interpretation of the results has been reliant on my hypotheses which propose relationships between token counts/D scores and the roles of the participating languages, whether that be as the Matrix or Embedded Language. Evidence has shown that the hypothesis about the D scores needs tweaking if it is to allow for the natural increase in lexical diversity resulting from linguistic development. However, despite this caveat, these hypotheses have proved to be extremely useful in enabling a thorough analysis of the VOCD results and have highlighted the code-switched data which would benefit most from further qualitative analysis. Before this

qualitative investigation begins, however, I will discuss the results gleaned from my final quantitative analyses which involve the command WDLEN.

4.3 WDLEN analyses and results

As described in the methodology chapter (section 3.3.6), the CLAN command WDLEN provides us with both the mean word length (MWL) and mean utterance length (MUL) of the specified input. It was pointed out that due to the specific language coding in the LOBILL Corpus, it would be possible to test two hypotheses which propose relationships between the results from WDLEN analyses and the relative participation of English and Portuguese in code-switched utterances. One hypothesis correlates a low MWL with the Matrix Language and a high MWL with the Embedded Language. The second hypothesis predicts that a low MUL will reflect the contribution of the Embedded Language while a high MUL will indicate a language acting as the Matrix Language. In this section the results of the WDLEN analyses will be examined in the light of these predictions. After presenting the results pertaining to mean word length, those relating to mean utterance lengths will then be discussed.

4.3.1 Mean Word Lengths (MWL) and code-switching

Before performing more specific WDLEN analyses which incorporated the variables of speaker, addressee and language, a very basic analysis was carried out on each of the speakers (12) in the corpus[65] resulting in 12 MWL scores (5 monolingual English scores, 3 monolingual Portuguese scores and 4 bilingual scores). The overall mean of the MWL values per groups of speakers proved to be very close: 3.70 for the English group, 3.89 for the Portuguese group and 3.75 for the bilingual group. A one-way ANOVA revealed that there was no significant difference between the groups ($F(2,9) = .734$, $p = .507$), reflecting what was found in the studies discussed in 3.3.6 - that there are no fundamental differences in word length between English and Portuguese. The fact that there is no significant variation across the three language groups is important as it will allow me to make valid comparisons across the WDLEN results as I investigate the relationship between MWL values and the Matrix/Embedded Language Asymmetry.

Having performed an initial basic WDLEN analysis, as for many of the FREQ and VOCD analyses, the next step was to specify the speaker, addressee and

[65] kwal @ +t*GRA +u +d | wdlen +r5 -s"@nonwords.cut"

language variables in each command line. And for the same reasons highlighted in the section on VOCD (4.2.2), I decided to restrict my analysis to the same eight speaker/interlocutor combinations. A fine-grained quantitative WDLEN analysis of PAI's very limited code-switching with his children and with his wife and that of MOT's with PAI would not be productive. Only a more qualitative approach is suitable when analysing the limited data produced in these particular speaker/interlocutor combinations.

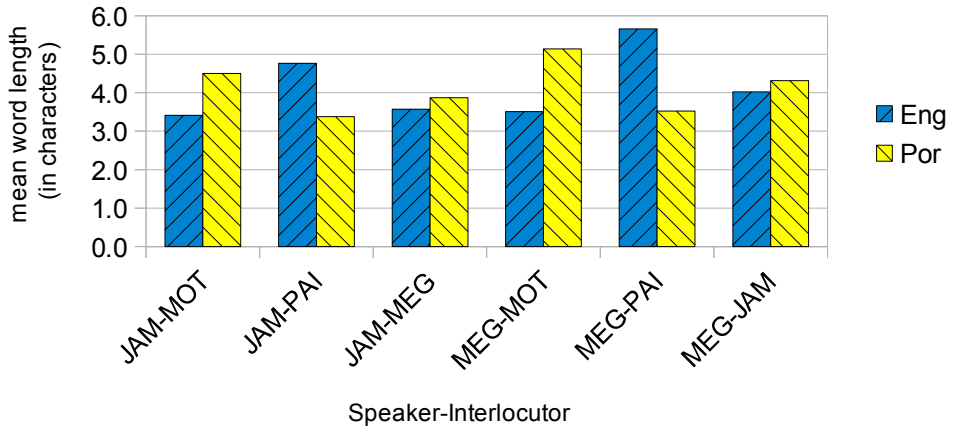
In order to investigate potential relationships between word lengths (in characters) and the roles of the two participating languages in CS utterances, two analyses for each of the eight speaker/addressee combinations were carried out. One provided the Mean Word Length of only English tokens in CS utterances[66] and the other provided the MWL of only Portuguese words found in CS utterances[67]. The results for JAM and MEG (six of the speaker/interlocutor combinations) will be discussed before I examine the MWL results pertaining to their mother.

4.3.1.1 MWL results for the siblings when code-switching with their parents and with each other

A first look at the chart in Fig. 19 may give the impression that there is no pattern to be found in the data for the siblings. However, if we were to transpose the first three clusters of columns (with JAM as speaker) over the second set of clusters (with MEG as speaker) we would find a close match.

[66]kwal @ +t%add +t*JAM +s"MOT" +u +d | wdlen +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@pt>"
[67]kwal @ +t%add +t*JAM +s"MOT" +u +d | wdlen +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@en>"

Figure 19. Mean word length (in characters) of English and Portuguese material in code-switched utterances for JAM and MEG per addressee



When addressing their mother, for both JAM and MEG the MWL for English is lower than that of Portuguese (3.41 as opposed to 4.49 for JAM and 3.51 as opposed to 5.11 for MEG. While the values for English fall slightly below the baseline of 3.74, the values for Portuguese appear to be significantly higher. According to my word length hypothesis, the differences in MWL values here indicate that with the mother both JAM and MEG are using English as the Matrix Language and Portuguese as the Embedded Language. When addressing their father, the opposite appears to be true: there are lower values for Portuguese (3.38 for JAM and 3.53 for MEG) and higher values for English (4.76 and 5.57), indicating the use of Portuguese as the ML and English as the EL. As for the MWL values when the siblings address each other, one observes that there is less disparity, the values for English (3.57 for JAM and 3.94 for MEG) being only a little lower than those for Portuguese (3.87 and 4.28 respectively). Neither language appears to be taking firm control of the CS utterances in this case.

As noted above, for all three sets of clusters the pattern of MWL values is found to be very similar for JAM and MEG. Such similarity in language role patterns was also found in the results of the FREQ and VOCD analyses (see 4.1.4 and 4.2.2.1) and therefore they add to the existing evidence that JAM and MEG's code-switching patterns (in terms of the role each language plays) appear to be mostly comparable if we control for the variable of addressee.

4.3.1.2 MWL results for the mother when code-switching with her children

When I first examined the WDLEN results for the remaining two combinations (MOT/JAM and MOT/MEG), I found MWL values that appeared to contradict my hypothesis. Instead of the predicted low MWL for English (which correlates with a Matrix Language) and a high MWL for Portuguese (correlating thus with an EL), the values were reversed: 3.67 for English and 2.72 for Portuguese when code-switching with JAM; 3.83 and 2.83 when code-switching with MEG. These values can be seen in (1) in Fig. 20.

The reader may recall that similar contradictory results had been found when performing VOCD analyses on exactly the same speaker/interlocutor combinations (see 4.2.2.2). In that case, through further investigation it was discovered that it was MOT's frequent use of 'o(lha)' (look) in CS utterances (49 occurrences when addressing JAM and 18 when addressing MEG) which had resulted in unexpectedly low D scores for Portuguese. By removing all tokens of 'o(lha)' from the VOCD analyses[68], the D scores had then fallen more in line with my expectations. It seemed reasonable to assume that MOT's idiosyncratic use of this Portuguese discourse marker may also be skewing the results of the WDLEN analyses in some way. I needed to examine the WDLEN output in more detail.

As 'olha' consists of 4 characters I expected to find a high frequency of 4-character words in both sets of results (at least 49 for the combination MOT/JAM and at least 18 for MOT/MEG). However, the frequencies found were low: 3 for MOT/JAM and 4 for MOT/MEG. It was actually under the 1-character word column that I found the high frequencies I was looking for: 50 and 18 respectively. WDLEN had clearly computed 'o(lha)' as a single character word. After consulting the CLAN manual (2013:122), I realised that my use of the +**r5** in the WDLEN command line had not only caused the command to ignore replacement material occurring in square brackets, but had also meant that any material (characters) in parentheses would also not have been counted[69]. As the vast majority of the 'olha' occurrences were transcribed as 'o(lha)', this had led WDLEN to count these tokens as single character words. Such high frequencies of single character words would clearly have resulted in a lower Mean Word Length.

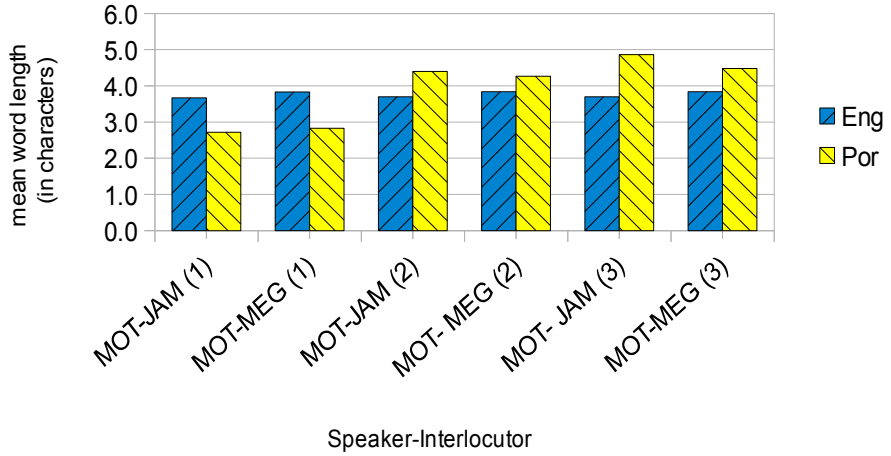
---

[68] Achieved with the following command line: Kwal @ +t%add +t*MOT +s"JAM" +u +d | vocd +r5 -s"@nonwords.cut" +s"[+ *]" -s"<@en>" -s"olha"

[69] It is only for the WDLEN command that the +**r5** switch has the effect of also excluding material in parentheses. For FREQ and VOCD, the +**r5** switch only has the effect of excluding replacement material found in square brackets.

The manual showed that it was possible to ask WDLEN to include bracketed material when computing word lengths by simply using the +**r1** switch instead of +**r5** . To see the effect that this method would have on the MWL values for both English and Portuguese I carried out the same four analyses but this time replaced +**r5** with +**r1**[70]. The output for Portuguese now showed high frequencies (51 and 21 respectively) appearing under the 4-character word column and this evidently led to a significant increase in the Mean Word Length for Portuguese, now 4.40 for MOT/JAM and 4.27 for MOT/MEG. The values for English showed very little change, increasing from only 3.67 to 3.70 for MOT/JAM and from 3.83 to 3.84 for MOT/MEG. These new values can be seen in (2) below.

Figure 20. Mean word length (in characters) of English and Portuguese material in code-switched utterances for MOT: with 'o(lha)' (1), with 'olha' (2) and without 'olha' (3)



Although the higher Portuguese MWL values are now more in line with the predictions of my hypothesis (i.e. reflecting the role of an Embedded Language), one could argue that by still including the frequent occurrences of 'olha' in the input, the results are still being skewed – just positively instead of negatively. As with the VOCD analyses, I decided to exclude all occurrences of 'olha' from the WDLEN input. This was achieved by adding the string -s"olha" to the command line along with the +**r1** switch[71]: the latter was needed to enable WDLEN to first identify the occurrences

[70]kwal @ +t%add +t*MOT +s"JAM" +u +d | wdlen +r1 -s"@nonwords.cut" +s"[+ *]" -s"<@pt>" and kwal @ +t %add +t*MOT +s"JAM" +u +d | wdlen +r1 -s"@nonwords.cut" +s"[+ *]" -s"<@en>"

[71]kwal @ +t%add +t*MOT +s"JAM" +u +d | wdlen +r1 -s"@nonwords.cut" +s"[+ *]" -s"<@en>" -s"olha"

of 'olha' before the former string instructed the programme to remove them. As can be seen in (3) in Fig. 20, the MWL values for Portuguese increased: to 4.86 for MOT/JAM and 4.48 for MOT/MEG. If we now compare the results in (3) (without 'olha') to those in (1) (with 'o(lha)') we see that for MOT/JAM the increase (the difference between the Portuguese MWL values) is 2.14 characters while for MOT/MEG it amounts to 1.65 characters. As exactly the same analyses were performed on both speaker/interlocutor combinations, these differences in MWL values can only be due to the difference in frequencies of 'olha': the removal of the 49 occurrences addressed to JAM had a greater effect on the resulting MWL value than the removal of those 18 addressed to MEG.

The above discussion has served to highlight the methodological challenges of using quantitative measures to analyse corpus data. If I had blindly accepted the MWL values provided in the output of the original analyses (see (1) in Fig. 20), I would have been led to conclude that my hypothesis regarding the relationships between low MWL/high MWL values and the Matrix/Embedded Languages was flawed. This was because previous FREQ analyses had already provided sufficient evidence for me to establish that when code-switching with her children MOT used English as the ML and Portuguese as the EL. Therefore, if, according to my hypothesis, the MWL values were showing a reversal of these roles, then one could only conclude that it was the hypothesis that was at fault. However, insights gained from the investigation of other seemingly contradictory evidence (the case of the unexpected D scores for MOT), led me to carry out a similar investigation here, and, as seen in the discussion above, it was found that it was MOT's idiosyncratic use of 'o(lha)' that was again responsible for skewing the results. There was no need to discard my MWL hypothesis.

Of course, by excluding all occurrences of 'olha' from MOT's analyses I am implying that they should not be counted as code-switched tokens. This is acceptable if we consider that in most cases MOT reduces the word to an unmarked 'o'. It may have made more sense to have classified such usage as belonging to the class of non-words (like 'ah' 'err'). However, if I had simply added 'o(lha)' to the **@nonwords.cut** file this would have meant that (i) all occurrences of the word (including more meaningful uses) would have been excluded, and that (ii) the exclusion would also apply to data pertaining to other speaker/interlocutor combinations. Although in other studies based on corpora (especially monolingual corpora) such methodological

choices might not have a major impact on the results, the nature of my investigation of code-switching means that such issues need careful consideration. This has been perfectly illustrated by the analyses discussed in this section.

Returning to my hypothesis, if we accept the third set of MWL values for the MOT/JAM and MOT/MEG combinations (see (3) in Fig.20), we have further evidence to support the idea that Mean Word Length values can be good predictors of the participatory roles of languages in CS utterances: for all eight speaker/interlocutor combinations, comparatively lower MWL values reflected a language being used as the Matrix Language while higher MWL values were seen to correlate with the Embedded Language. I will now turn to my second WDLEN hypothesis and show how the measure of Mean Utterance Length can be used to further enhance the investigation of code-switching in my study.

4.3.2 Mean Utterance Lengths (MUL) and code-switching

When talking about the results in this section it is important to point out that for the most part, the mean utterance length (MUL) values actually refer to the relevant parts of each code-switched utterance. Therefore, the Portuguese MUL for any given speaker/addresse combination is the average length (in words) of only the Portuguese contribution to the CS utterances and the English MUL refers to the average number of words contributed by English. An existing measure traditionally used to investigate child language development is being applied in a novel way here. But this is only possible due to the specific language coding of the LOBILL Corpus which allows WDLEN to separate the two languages found in CS utterances.

It was not productive to compare the overall MUL values across the speakers in the LOBILL Corpus as the variation found merely appeared to reflect the diversity you would expect in spoken discourse in terms of utterance length. However, it did prove insightful to make a comparison between the MUL values of monolingual and bilingual utterances for the eight bilingual speaker-interlocutor combinations. These analyses are discussed below.

4.3.2.1 A comparison of MUL values of monolingual and bilingual utterances

In order to compare the MUL values of these two modes of speech, two WDLEN analyses were performed on each of the 8 speaker/interlocutor combination: the first selected only monolingual utterances for analysis (the -s"[+ *]" switch removing all

CS utterances from the input)[72] and the second selected only CS utterances (+s"[+ *]")[73]. It is important to point out that although the monolingual input contained both English and Portuguese data, what I am interested in here is the contrast between monolingual utterances (i.e. where no code-switching takes place) and those utterances in which both languages participate. The results can be seen in Table 10 below:

Table 10. Mean Utterance Length (MUL) of monolingual utterances (English and Portuguese combined) and MUL of only CS utterances per speaker/interlocutor combination.

Speaker-interlocutor	MUL of monolingual utterances	MUL of CS utterances
JAM - MOT	3.66	7.18
JAM - PAI	4.38	9.11
JAM - MEG	3.36	5.79
MEG - MOT	4.31	7.39
MEG - PAI	5.59	11.84
MEG - JAM	4.20	5.31
MOT - JAM	4.69	5.93
MOT - MEG	4.48	4.93

By comparing both values what we find is that in every single case, without exception, the Mean Utterance Length for CS utterances is higher than that for monolingual utterances. The disparity in values may, at times, be slight, such as for the combinations MEG/JAM, MOT/JAM and MOT-MEG where the monolingual MUL values of 4.20, 4.69 and 4.48 respectively are slightly below that of the corresponding CS MUL values 5.31, 5.93 and 4.93. However, we can also see larger differences, such as for MEG/PAI where her MUL for monolingual utterances of 5.59 is over 5 words less than the MUL for CS utterances, which is 11.84. A paired t-test revealed that the differences between the MULs of monolingual utterances and of CS utterances were indeed significant (t=-4.095, df=7, p=.005) and confirmed that, in the LOBILL Corpus at least, a CS utterance is typically longer in words than the average monolingual utterance, independent of who is speaking or being spoken to. Is this a typical characteristic of CS discourse? Further analyses (discussed in 5.2.2.3 and

[72] kwal @ +t%add +t*JAM +s"MEG" +u +d | wdlen +r5 -s"@nonwords.cut" -s"[+ *]"
[73] kwal @ +t%add +t*JAM +s"MEG" +u +d | wdlen +r5 -s"@nonwords.cut" +s"[+ *]"

6.3) will shed more light on these particular quantitative results. If the same WDLEN analyses were performed on other bilingual corpora and revealed the same findings, this would indeed indicate that CS utterances are characterized as being relatively longer than monolingual utterances. Of course, decisions made at the time of transcribing are fundamental in determining the outcome of such results: only consistency in terms of determining utterance boundaries will lead to reliable results.

Before leaving the discussion of Table 10, it is worth pointing out a similarity which can be found between the siblings in terms of the MUL values and their interlocutors. For both JAM and MEG, the MUL values (monolingual and CS) when addressing PAI are higher than those for MOT (compare the CS MUL of 9.11 for JAM and 11.84 for MEG when addressing PAI to 7.18 and 7.39 when interacting with MOT). And the lowest CS MUL values for JAM and MEG (5.79 and 5.31) are when they are interacting with each other. Again we find evidence to suggest similarities in the siblings' code-switching practice: their MUL values are comparable in terms of addressee.

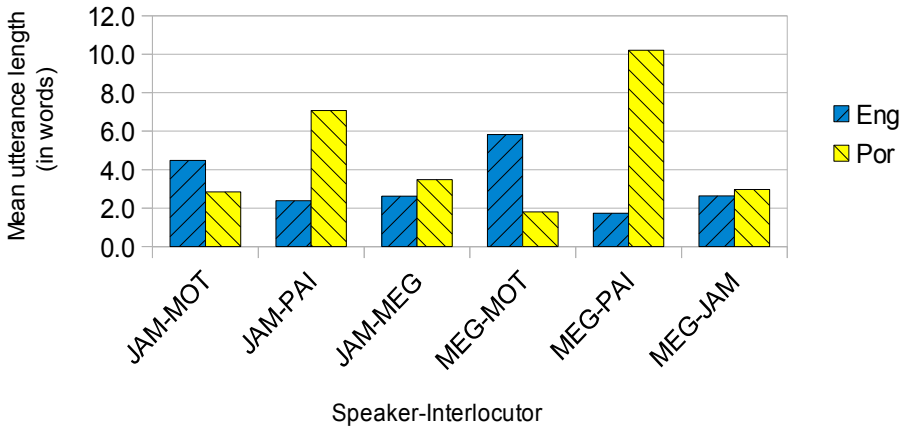
It is when we examine the MUL values for the English and Portuguese contributions to CS utterances that we are able to learn more about their participatory roles in terms of the ML/EL asymmetry. I will first discuss the results pertaining to JAM and MEG (six of the speaker/interlocutor combinations) before looking at those relating to MOT (MOT/JAM and MOT/MEG)

4.3.2.2 MUL results for the siblings when code-switching with their parents and with each other

The command lines used to perform the analyses discussed in this and the next section were actually the same as those which were used to output the Mean Word Lengths in 4.3.1 (see footnotes 66 and 67 for the respective command lines). This is because both MWL and MUL values can be found in the same output. Therefore, no new analyses were necessary.

As can be seen in Fig. 21 the results show that JAM and MEG again share a similar pattern in terms of their MUL values per language per addressee.

Figure 21. Mean Utterance Length (in words) of English and Portuguese material in code-switched utterances for JAM and MEG when addressing MOT, PAI and each other.



When the addressee is their mother, English has a higher MUL than Portuguese and when it is their father the opposite occurs. When addressing each other, the MUL values are relatively close. Despite the similar patterns, it is evident that MEG's values are more disparate than those for JAM when the addressees are the parents. With MOT as interlocutor, JAM's MUL scores are 4.48 for English and 2.85 for Portuguese. That is, in terms of their contribution to CS utterances, English accounts for approximately 61% of the utterance, the remaining 39% being proportioned by Portuguese. The percentages for MEG reveal a wider disparity in values: 76% (with an MUL value of 5.82) being contributed by English and 24% (with an MUL of 1.83) by Portuguese. Such relative disparity is also evident when we examine the values for PAI as addressee: the proportion of English to Portuguese for JAM is 25%/75% (MULs of 2.38/7.07) while for MEG it is 14%/86% (MULs of 1.72/10.28). Yet again, according to the MUL hypothesis, these results are indicative of a more classic style of code-switching employed by MEG when compared with her brother.

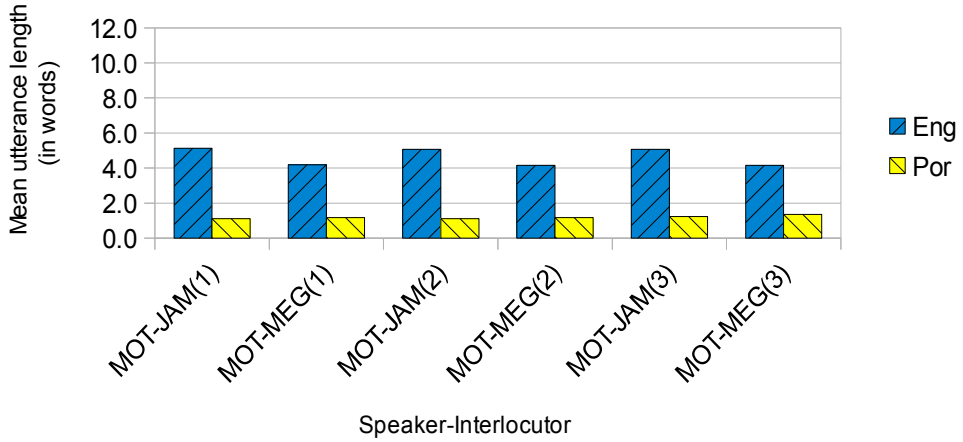
It is not a coincidence that the percentages mentioned above almost exactly reflect the percentages discussed in the frequency results (see section 4.1.4). For this study WDLEN is being used to investigate the relative contribution of English and Portuguese tokens to CS utterances and the output is given as the mean number of tokens for each language per utterance. If we were to take a speaker's total token counts for English and Portuguese in CS utterances (provided by FREQ) and divide these by the number of CS utterances, we would arrive at the average number of

tokens each language contributes to each CS utterance. Thus, the end results would theoretically be the same. However, this methodological insight was only made possible after both sets of results were converted (manually) to relative percentages and compared. Of course, the WDLEN analyses provide us with additional data on word and utterance length and these results cannot be provided by FREQ. And in the same turn FREQ is not just about token counts but is able to provide word lists and even concordances. However, what I specifically wish to highlight in the discussion of WDLEN is how an existing measure, such as WDLEN, can be exploited methodologically in order to provide original results.

4.3.2.3 MUL results for the mother when code-switching with her children

Returning to the results, let us now see whether the MUL values of MOT reflect what the MUL hypothesis predicts about the relationship between MUL values and the participatory roles of English and Portuguese in CS utterances. For reasons of consistency, I have decided to show the three sets of results which correspond to the three different analyses where (1) 'o(lha)' was counted as a single character word, (2) 'olha' was counted as a four character word and (3) all occurrences of 'olha' were removed from the input. As mentioned previously, there was no need to carrry out new analyses.

Figure 22. Mean utterance length (in words) of English and Portuguese material in code-switched utterances for MOT when addressing JAM and MEG: with 'o(lha)(1), with 'olha'(2) and without 'olha'(3)



The first thing to point out in Fig. 22 is that there is consistency across the three types of results in terms of MUL values. Whereas the results in Fig. 20 revealed that the MWL values were significantly affected by changing the input (especially when all occurrences of 'olha' were excluded from the data), here there is little, if any, difference when we compare the MUL values across (1), (2) and (3).

Looking first at when MOT addresses JAM, we see noticeably higher MUL values for English than for Portuguese: 5.12 (1), 5.06 (2) and 5.06 (3) as opposed to 1.11 (1), 1.11 (2) and 1.23 (3). This means that when code-switching with her son, MOT's utterances consist of an average of 5 English words and 1 Portuguese word - this reveals a clear asymmetry in the use of both languages. According to my hypothesis such values reflect the fact that English (by contributing more words) is acting as the Matrix Language and Portuguese (by contributing relatively fewer words) is taking on the role of the Embedded Language. A similar pattern can be found when MOT code-switches with her daughter. With MEG as addressee, the MUL values for English are 4.18 (1), 4.15 (2) and (4.15) and for Portuguese they are 1.17 (1), 1.17 (2) and 1.35 (3). Rounding down the numbers, the average number of words contributed by each language to a typical CS utterance is 4 for English and 1 for Portuguese. These MUL values are again evidence of MOT's asymmetrical use of both languages when code-switching with her children and further support the similar findings resulting from the FREQ analyses in 4.1.4.

With regards to the very slight differences in MUL values found across the three different analyses (just visually perceptible in Fig. 22), the slight increase seen for Portuguese can be simply explained. One would expect the Portuguese MUL values for (1) and (2) to be the same since both forms of 'o(lha)' (whether consisting of 1 or 4 characters) are being counted as single words. Indeed the results show that this the case. However, by removing all occurrences of 'olha' (3), the utterances where this word represented the only Portuguese contribution, are effectively excluded from the MUL calculation resulting thereby in an increase in the Portuguese MUL. The fact that the increase noted is so slight must mean, however, that most of the contribution of Portuguese to CS utterances is in the form of single words with only some occurrences of two words or more (such as in an Embedded Language Island).

With regards to the English MUL values, for each addressee we see exactly the same values for analyses (2) and (3): 5.06 for MOT/JAM and 4.15 for MOT/MEG.

This is to be expected since the removal of 'olha', a Portuguese word, would not effect the English results. However, there is a very slight increase in MOT's MUL values for (1): 5.12 when addressing JAM and 4.18 when the addressee is MEG. Clearly this increase must be due to the use of the +**r5** switch in analysis (1) as opposed to the +**r1** switch in analyses (2) and (3) as this is the only difference between the analyses. Within the confines of this study it is not feasible to investigate further as this would mean searching the transcripts for tokens where the inclusion/exclusion of material in brackets would have had such an effect. In any case, such minimal variation in the English results (increases of only 0.06 for MOT/JAM and 0.03 for MOT/MEG) do not have an impact on my claims about what MUL values can reveal about the participatory roles of languages in CS utterances. As seen in the discussions in this section and in 4.3.2.2 above, there are clear relationships to be found between Mean Utterance Length values and bilingual language use in code-switched discourse: high MUL values are indicative of a Matrix Language while comparatively low MUL values characterise the contribution of an Embedded Language.

In this study, we have seen how the command WDLEN can be used to investigate the participatory roles of English and Portuguese in the CS utterances found in the LOBILL Corpus. Such use of WDLEN, as demonstrated here, is believed to be original in the sense that it offers a novel approach to the quantification and interpretation of the asymmetry principle, a characteristic of classic code-switching. The two hypotheses posed at the beginning of sections 4.3.1 and 4.3.2 have allowed for a fruitful analysis of the results and the evidence appears to support what they propose: that there are strong relationships between both Mean Word Length values and Mean Utterance Length values and the roles of the participating languages in CS utterances: a combination of low Mean Word Length/high Mean Utterance Length values reflect the role of the Matrix Language while a combination of high Mean Word Length/low Mean Utterance Length values reflects that of the Embedded Language. Where the values are less disparate, the two languages could be said to be participating more equally, both in terms of types of tokens being contributed (measured by MWL) and numbers of tokens contributed to code-switched utterances (measured by MUL).

The quantitative investigation of code-switching reported on in this chapter has demonstrated ways in which traditional quantitative measures can be exploited in order to learn more about structural aspects of the participating languages of code-switched speech. Not only was it possible to confirm the existence of a Matrix/Embedded Language asymmetry in the data in terms of the word frequency of each contributing language, the VOCD and WDLEN results provided evidence to suppport my own hypotheses regarding relationships between the ML/EL and three other quantitative measures (vocabulary diversity, word length and utterance length). These four types of measures combined offer researchers a novel way of establishing the roles each language has to play in an individual's bilingual utterances. Based on my own results, I have developed a simple schema which aims to aid researchers wishing to interpret the different values arising from the use of such measures in their own code-switched data. This schema will be presented and discussed in Chapter 8. Presently, however, I will now turn to the second level of analysis performed on the code-switched material in the LOBILL Corpus, a word and code level analysis.

5. Word and code level analyses and results

In Chapter 4 the focus of discussion was the quantitative output of the analyses performed on the data using the commands FREQ, VOCD and WDLEN. The resulting numerical values were analysed and associations were made between the four different types of values and what they indicate in terms of the Matrix and Embedded Languages of code-switched utterances. In this chapter the aim is to take the investigtion of code-switching in the corpus to a different level of analysis, one which involves a word-based and code-based approach to the data. That is to say, the focus here will be on the examination and interpretation of word lists and code lists, mostly produced by the use of the command FREQ. Although quantification remains a key aspect of these analyses, we are now more interested in what the output can tell us about the nature of the code-switched utterances. As will be seen, due to the insertion of different types of codes in the corpus it was possible to perform several types of analyses on the data, all of which reveal interesting aspects about the code-switching employed by the main informants. As for the previous sections, all command lines used will be exemplified in the footnotes.

5.1 Frequency word lists of code-switched material

In section 4.1 the FREQ command was used to perform increasingly more specific analyses of the data in order to calculate total numbers of tokens. At their most specific they were designed to incorporate the variables of speaker, addressee and language (the proportion of English and Portuguese words contributing to code-switched utterances). The resulting token counts indicated the roles each contributing language had to play when the siblings and their parents engaged in code-switching with each other: a relatively higher proportion of words indicated a language acting as the Matrix Language whereas a lower proportion was indicative of the Embedded Language. In this section we turn from numbers of tokens to the actual words themselves and see what the word lists reveal about the nature of each language's contribution to CS utterances.

Previous FREQ analyses had shown that JAM and MEG engaged in significantly more code-switching than their parents (see Figs. 6 and 7). Higher CS token counts for the siblings meant that it was possible to perform more specific analyses on their utterances than it was on either MOT's or PAI's CS utterances. For

example, only for the siblings was it possible to investigate differences in lexical diversity across the seven interaction types (see Figs. 13 and 14). Due to lower CS token counts for the parents, such specificity would not have produced any results. Such is the case for the FREQ analyses reported on in this section. Based on the data in Fig. 7, the only speaker/addressee combinations that would allow for the incorporation of the variable of interaction type were those that involved the siblings addressing their mother (that is, JAM/MOT and MEG/MOT). With regards to the siblings' interactions with their father, it was only possible to perform the analysis on one interaction type, that of the telephone conversations (TI). Although in the latter case no comparison could therefore be made across the interaction types, this analysis did prove to be productive, as will be seen in the discussion below. For most of the other speaker/addressee combinations (MOT/JAM, MOT/MEG, JAM/MEG and MEG/JAM) the analyses were performed on all the files at once, thereby merging the tokens from all the interaction types. It was also decided that due to the very low numbers of CS tokens for PAI the FREQ analyses would not be carried out on his utterances; these would only be examined at utterance, and not word, level.

Focussing first on JAM and MEG's code-switching with their mother, two steps were necessary to carry out the FREQ analyses. Firstly, all the files for one of the interaction types were selected (via the "File in" button in the commands window). Secondly KWAL and FREQ were used to perform the analyses in the following way: KWAL was used to select all the utterances pertaining to a specific speaker/addressee combination and then FREQ was used to select only the code-switched utterances and provide separate word lists for each contributing language[74]. This gave rise to a total of 28 analyses where the same two basic command lines systematically went through a process of speaker and language substitution (2 speakers x 1 addressee x 2 language types x 7 interaction types). When all the analyses for one interaction type were complete, these files were substituted by those of another interaction type and exactly the same analyses were performed again, and so on. For those other speaker/addressee combinations where the CS tokens from the different interaction types needed to be merged (a further 6 analyses), step one (the selection of files according to interaction type) was missed out. Instead, the 'Add All' button in the commands window ensured that all the files in

[74]kwal @ +t%add +t*JAM +s"MOT" +u +d | freq +r5 +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +o and kwal @ +t%add +t*JAM +s"MOT" +u +d | freq +r5 +s"[+ *]" -s"@nonwords.cut" -s"<@en>" +o

the corpus were included for analysis. Apart from this, the command lines themselves remained the same as for the first 28 analyses.

The resulting word lists were then scrutinized and interpreted in terms of the ML/EL asymmetry. Due to textual restrictions the word lists used to illustrate my interpretation of the results have been truncated to show just the top 20 most frequent words appearing in the lists[75]. To faciliate comparison the total numbers of types and total tokens for each word list is given at the bottom of the columns and, in addition, translations have been supplied for the Portuguese word lists (note that *-m* is masculine and *–f* is feminine).

In order to interpret the word lists in terms of the ML/EL asymmetry, it is first important to recall Myers-Scotton's 4-M Model which offers a classification of morpheme types based on whether they are conceptually activated or structurally assigned (see section 2.1.1.1 for more detail). For the purposes of this discussion what needs to be highlighted here is that in classic code-switching the Matrix Language typically contributes the morphemes which make up the grammatical framework of the utterance (such as early system, bridge and outsider morphemes) while the Embedded Language typically contributes content morphemes, examples of which are nouns, verbs, and adjectives. By comparing the word lists for each language in terms of morpheme types it should be possible to determine whether this Matrix/Embedded Language distinction is evident. Although the classification of some of the words may be challenging because they are out of their original linguistic context, this did not prevent conclusions being drawn from the data.

5.1.1 Frequency word lists of the siblings when code-switching with their mother in Meal Time (MT) interactions

The first set of results I will discuss are the word lists resulting from the analyses of JAM and MEG's interactions with their mother at meal times[76]. Earlier FREQ, VOCD and WDLEN analyses had identified that both siblings appeared to use English as the ML and Portuguese as the EL when code-switching with MOT. This finding is supported by an analysis of the word lists in Table 11. For both JAM and MEG, among the top 20 most frequent English words there are articles, possessive

[75] To allow for more effective comparison, all proper names (e.g. of people and places) were removed from the lists (manually), excepting those where there was a translation equivalent (e.g. England/Inglaterra). Spanish words were also excluded.

[76] Files 13-15, 17, 20, 21, 23, 35, 39, 50, 73, 76, 79-81, 83, 84, 86-90, 97-99, 108 and 111.

adjectives, personal pronouns, prepositions, conjunctions and a relative pronoun, all of which are types of system morphemes typically contributed by the ML. Although it is not possible to determine how many of the occurrences of 'to' (23 for JAM and 15 for MEG) and 'of' (13 occurrences for JAM) should be assigned as early system morphemes or late system morphemes (see discussion in Myers-Scotton, 2002:79-81), here the fact that they are not content morphemes is significant enough. The lack of nouns, adjectives and verbs conveying semantic content in the word list for English is noticeable and does indeed indicate that English is taking on the role of the Matrix Language by supplying the grammatical framework for CS utterances. In addition, the fact that JAM and MEG share 12 of the top 20 words in the English lists is clear evidence that English is performing an extremely similar role, that of the ML, in the CS utterances of both siblings.

Table 11. Frequency word lists per language for JAM's and MEG's CS utterances when addressing MOT in Meal Time interactions.

JAM-MOT		MEG-MOT	
English	Portuguese	English	Portuguese
71 the	12 de (of)	47 the	6 de (of)
48 and	10 eu (I)	21 I	5 ensaio (rehearsal)
27 is	7 está (is/okay)	19 it	5 flocos (chocolate chips)
24 I	7 leão (lion)	18 and	4 burro (donkey)
23 to	6 buraco (hole)	18 you	4 zero (zero)
21 that	6 e (and)	15 to	3 centro (centre)
16 just	6 em (in)	14 that	3 direito (right)
16 my	6 vermelho (red)	11 he	3 esquerda (left)
15 a	6 é (is/yes)	10 a	3 filha (daughter)
13 of	5 areia (sand)	9 but	3 geral (general)
12 but	5 coleguinha (classmate)	9 don't	3 vinte (twenty)
12 going	5 gorilla (gorilla)	9 this	3 é (is/yes)
12 he's	5 quando (when)	8 is	2 India (India)
12 you	5 robô (robot)	8 on	2 a (the - f)
11 his	5 um (a/one - m)	8 or	2 atolou (got stuck)
11 it	4 cima (above)	8 was	2 aventureiros (adventurers)
11 look	4 papapa (food)	7 because	2 cursiva (joined up)
11 on	4 tu (you)	7 do	2 dança (dance - n)
11 one	4 uma (a/one – f)	7 in	2 e (and)
11 there	4 vai (go/goes)	7 like	2 ele (he)
Types: 184 Tokens: 824	Types: 225 Tokens: 401	Types: 216 Tokens: 638	Types: 158 Tokens: 213

If we now examine the lists of Portuguese words for JAM and MEG we are able to note several differences compared to the English word lists. Before commenting on the distribution of morpheme types there are two important observations to be made. One can observe first of all that the frequency of each of the top 20 words contrasts quite dramatically with that of the English words. These much reduced numbers are indicative of the smaller role the Embedded Language has to play in terms of overall contribution to CS utterances. The next thing to observe is that there also appears to be a contrast in terms of word length: with the exception of a few words (de, e, é and a), the words in the Portuguese lists are noticeably longer (in characters) than those in the English word lists. This corroborates what the WDLEN analyses had shown regarding relationships between mean word length and the ML/EL, namely that a relatively longer mean word length appeared to be indicative of a language acting as the Embedded Language in CS utterances. This is because content words tend to be longer in length when compared with grammatical words. And, in fact, if we now examine the distribution of morpheme types in the Portuguese word lists we find that in the top 20 words for both JAM and MEG there are several nouns: seven in JAM's list (leão, buraco, areia, coleguinha, gorilla, robô and papapa) and eight in MEG's list (ensaio, flocos, burro, centro, filha, India, aventureiros and dança). All of these are at least four characters in length. Other types of content morphemes appearing in the lists are an adjective (vermelho) for JAM and cursiva, geral, direito and esquerda for MEG), a verb (atolou for MEG), and a preposition (cima for JAM). The presence of such content words clearly points to the fact that Portuguese is being used as the Embedded Language by JAM and MEG when code-switching with their mother. Furthermore, in contrast to the English word lists, where JAM and MEG shared 12 types of system morphemes, there is no similarity at all in their use of content words. It will be important to see if this pattern is repeated in the results of the other interaction types.

Whereas the content words mentioned above are exclusive to the Portuguese word lists (i.e their translation equivalents do not appear in the top 20 words of the English lists), it is interesting to note that there *are* items which do appear in both lists. In both JAM and MEG's lists we find the conjunction and (appearing 48 times in JAM's lists and 18 times in MEG's list) and its Portuguese equivalent e (6 and 2 occurrences respectively). Another item which can be found in both language lists is the word is which can be translated by the Portuguese é. However, in Portuguese,

the latter is frequently used in the sense of 'yes (it is)' as a stand-alone short answer. Only by examining each case in its wider linguistic context will further interpretation be possible. Another word which appears in both of MEG's lists is he (11 occurrences) and its Portuguese equivalent ele (2 occurrences), an early system morpheme which, according to the 4-M model, can be provided by either the ML or the EL. Why MEG uses the Portuguese pronoun instead of the English equivalent on two occasions can only be investigated by looking at the utterances in which they occur. Finally, in JAM's lists we find the word of (13 occurrences) and de (12 occurrences), the latter ranking as the most frequent Portuguese word in the CS utterances of both JAM and MEG (6 occurrences). This particular finding merits immediate discussion and will briefly be commented on below.

Myers-Scotton classes both the English of and the genitive -'s as bridge-system morphemes which are typically contributed by the Matrix Language in CS utterances (2002:79). It is therefore perhaps suprising to find that the Portuguese de appears to be the most frequent contribution from the Embedded Language! Even if we consider that the only way to express genitive case in Portuguese is through de, thereby increasing its frequency when compared to English, this still does not explain why this system morpheme (untypical of an Embedded Language contribution) should be found so high up in the Portuguese word lists. One explanation I would like to put forward which would account for this apparent contradiction lies in the fact that de also functions as a compound noun linker. For example, the compound noun 'dining table' would be realized in Portuguese by 'mesa de jantar' and the equivalent of 'sunglasses' is 'óculos de sol'. These are clearly lexical units, the 'de' forming part of compound noun and, as a result, should not be counted as separate system morphemes. It is plausible, therefore, that many of the cases of de in the Portuguese lists are performing this function and are not actually bridge-system morphemes at all.[77] Only an utterance level analysis would reveal how the de is being used in each case.

Apart from comparing the frequency of content/system morphemes in the word lists resulting from the analyses of the other interaction types, it will also be

[77]Within the CHAT transcription system, compounds can be transcribed using linking symbols; the Portuguese examples could therefore be transcribed as '*mesa+de+jantar*' and '*oculos+de+sol*'. By doing so, this type of *de* is not counted as a separate morpheme by FREQ. Although with hindsight this would have been the preferred method of transcribing compounds in Portuguese, the implications of the non use of the linkers in these cases was not apparent at the time of the compilation and transcription of the LOBILL Corpus. This would be a recommendation for the future.

interesting to see if the pairs mentioned above (and/e, is/é, he/ele and of/de) occur in JAM and MEG's lists and, if they do, with what frequency. Such occurrence would then merit further investigation at utterance level (see section 6.2). Before doing this, however, I will discuss what can be gleaned from the frequency word lists which resulted from the FREQ analyses of JAM and MEG's telephone interactions with their father.

5.1.2 Frequency word lists of the siblings when code-switching with their father in Telephone Interactions (TI)[78]

It became apparent when performing the VOCD analyses (in 4.2.2) that the only interaction type where the siblings addressed their father (PAI) with significant frequency was when they talked to him over the phone while they were in England (on holiday and after having moved there): the total number of utterances addressed to PAI amounted to 1491[79]. Previous FREQ, VOCD and WDLEN analyses had shown that when code-switching with their father, both JAM and MEG appeared to use Portuguese as the Matrix Language and English as the Embedded Language. An examination of the word lists below (Table 12) will determine whether this ML/EL asymmetry is bourne out in terms of morpheme types[80].

Looking first at the Portuguese word lists, what we find are system, or grammatical, morpheme types and a noticeable lack of semantically-laden content morphemes. The word length contrast observed in the lists in Table 10 is also visually apparent: with the exception of five words for JAM and four for MEG, all the Portuguese words are a maximum of three characters in length, contrasting with the longer English words. If we also consider that 15 of the 20 Portuguese words are exactly the same for JAM and MEG, this is ample evidence to support the hypothesis that both siblings are using Portuguese as the Matrix Language when code-switching with their father, and, must be doing so in an extremely similar way.

Table 12. Frequency word lists per language for JAM and MEG's CS utterances when addressing PAI in Telephone Interactions.

JAM-PAI		MEG-PAI	
English	Portuguese	English	Portuguese

[78] Files 59, 60, 62, 64, 65, 69, 71, 72, 74, 75, 93-95, 100, 102, 104, 106, 109, 110, 114-119.

[79] This number was gleaned from the output of the following FREQ analysis: freq @ +t%add -t* +u +o performed on only the TI group of files.

[80] See footnote 74 for the command line used but note that "MOT" was substituted by "PAI".

9 train	47 um (a/one -m)	7 and	79 e (and)
6 tram	35 que (that/who)	4 library	49 que (that/who)
4 and	35 é (is/yes)	4 triceratops	42 um (a/one -m)
4 bath	30 e (and)	3 dictionary	39 o (the -m)
4 but	26 eu (I)	3 guinea+pigs	33 de (of)
4 live	25 o (the -m)	3 no	33 eu (I)
4 train+track	24 a (the -f)	3 rock	26 a (the -f)
3 because	21 não (no)	3 tram	26 tem (has/there is)
3 digger	19 tem (has/there is)	2 French	22 no (in the -m)
3 in	17 mas (but)	2 black	20 lá (there)
3 lightening	16 gente (we/people)	2 field	19 aí (so)
3 seaside	14 no (in the -m)	2 kittens	17 é (is/yes)
3 swimming+pool	12 só (only/alone)	2 name	15 para (to/for)
3 van	12 também (also)	2 oink	13 tinha (had/there was)
2 England	9 aí (so)	2 p@l	12 estava (was)
2 all	9 quando (when)	2 pigeon	12 gente (we/people)
2 boyfriend	8 ele (he)	2 river	12 na (in the -f)
2 crazy	8 eles (they)	2 sports+day	10 não (no)
2 don't	7 estava (was)	2 squash	9 ele (he)
2 her	7 para (to/for)	2 station	9 pro (to the -m)
Types: 85 Tokens:149	Types: 171 Tokens: 670	Types: 127 Tokens: 174	Types: 300 Tokens: 1012

With regards to the role of English, we find a prevelance of content words in the lists: in JAM's list there are 11 nouns, one verb and one adjective; in MEG's list there are 14 nouns and two adjectives. There can be no doubt that here English is contributing with typical Embedded Language items. Amongst the remaining English words we find conjunctions: and occurs four times in JAM's list and tops MEG's frequency list with 7 occurrences; but and because appear in JAM's list four and three times respectively.

As was the case for the previous word lists the conjunction and/e makes a noticeable appearance in both language lists, the Portuguese equivalent ocurring 30 times in JAM's list and 79 times in MEG's list! It will be interesting to see whether such a high frequency of this conjunction is to be found in the other five interactions types. Three other translation equivalents appear in JAM's lists: the conjunction but/mas (4/17 occurrences), the preposition in/no (3/14 occurrences and the negative don't/nâo (2/2 occurences). In MEG's lists there is only one other pair, no/nâo with *3/10* occurrences respectively. At this point it would be pure speculation to suggest why both JAM and MEG would sometimes use the English equivalent when code-switching with their father when the Portuguese equivalent was clearly in frequent

use (no need to fill a lexical gap for example). However, through the examination of the utterances (in section 6.2) the reasons behind such usage will become apparent.

From the examination of the lists above (from Meal Time and Telephone Interactions) we have found more evidence to strengthen the claim that there is clearly an ML/EL asymmetry at work in JAM and MEG's code-switched utterances when addressing their mother and father. By applying Myers-Scotton 4-M model to the data it has been possible to show the following: when the siblings address their mother in meal time interactions, English plays the role of Matrix Language while Portuguese contributes as the Embedded Language; when they address their father over the telephone the reverse is true. With regards to the other interaction types it will only be viable to examine the CS utterances addressed to MOT as PAI's participation does not provide sufficient data for an effective word-level analysis. Furthermore, due to textual constraints, I will not be able to present such detailed interpretations of the word lists pertaining to the remaining interaction types. Instead, I will summarise the findings in the form of a table, focussing on the presence/absence of content words in the siblings' word lists for each language: their presence in the lists would indicate a language acting as the EL and their relative absence would indicate a Matrix Language.

5.1.3 The 4-M model applied to the word frequency lists of the siblings when code-switching with their mother in other interaction types

It became apparent in the discussions above that the classification of the words from the lists into one of the four morpheme types can be problematic. This is because we are examining words out of their linguistic context. To enable a more reliable comparison across interaction types and between the speakers JAM and MEG it was therefore decided to only include in the content word count those words which were nouns (N), verbs (V) (excluding modal verbs, copula forms and auxilaries) and adjectives (A). Each type was counted only once, irrespective of its frequency, and therefore the totals shown below in Table 13 (in columns five and seven) are out of 20 (most frequent types). Column three shows the total number of English and Portuguese tokens for each interaction type and was included so the possible effect of differences in token size could be taken into account. The discussion that follows will compare the siblings' use of content words in both languages in the seven

different interaction types and offer an interpretation of what the patterns indicate about the ML/EL asymmetry in their code-switching with their mother.

Table 13. Number of English and Portuguese content words (types) per interaction type occurring in the CS utterances of JAM and MEG addressed to MOT (top 20 occurrences)

Interaction type (No. of files)	Speaker	Total English/ Portuguese tokens	English content words in CS utterances		Portuguese content words in CS utterances	
			N V A*	Total	N V A	Total
MT (27)	JAM	824/401	0 0 0	0	7, 0, 1	8
	MEG	638/215	0 0 0	0	8, 1, 4	13
CH (20)	JAM	723/498	0 0 0	0	3, 1, 0	4
	MEG	217/86	1, 2, 0	3	11, 4, 1	16
PG (15)	JAM	152/177	1, 1, 1	3	0, 3, 2	5
	MEG	163/83	1, 0, 0	1	6, 3, 3	12
TI (25)	JAM	54/35	2, 3, 0	5	5, 7, 0	12
	MEG	41/11	1, 6, 0	7	4, 1, 0	5
LA (11)	JAM	26/43	2, 1, 1	4	3, 4, 1	8
	MEG	188/20	1, 1, 1	3	10, 5, 0	15
FP (14)	JAM	389/244	1, 1, 0	2	2, 4, 0	6
	MEG	124/43	0, 0, 0	0	5, 3, 3	11
IN (6)	JAM	36/27	4, 2, 0	6	4, 2, 0	6
	MEG	189/36	1, 4, 0	5	7, 5, 0	12
All files	JAM	2204/1425	10, 8, 2	20	24, 21, 4	49
	MEG	1560/494	5, 13, 1	19	51, 22, 11	84

* N = noun, V = verb and A = adjective

If we begin by comparing the total number of content words (types out of 20) per interaction type across both languages (columns five and seven) what we find is that the totals for Portuguese (column seven) are consistently higher than that for English in all but two cases. Leaving the latter exceptions aside for the moment, what these higher totals are telling us is that Portuguese is contributing more content words than English and is therefore likely to be acting as the Embedded Language. The fact that MEG's totals for Portuguese reach double figures in six out of seven of the interaction types (13 for MT, 16 for CH, 12 for PG, 15 for LA, 11 for FP and 12 for IN) is a strong

indication of how typically 'classic' her use of the Embedded Language is (by contributing a high proportion of content words). For the same interaction types JAM's totals are lower: 8 for MT, 4 for CH, 5 for PG, 8 for LA, 6 for FP and 6 for IN. A lower proportion of content words means a higher proportion of system morphemes, a contribution less typical of the Embedded Language. This is further evidence to support earlier findings (from the FREQ, VOCD and WDLEN analyses) that JAM's CS patterns seem to be less 'classic' than his sister's.

There are two exceptions where Portuguese does not contribute more content words than English in the siblings' CS utterances addressed to MOT. The first of these can be seen in MEG's totals for TI where there are 5 Portuguese content words as opposed to 7 English content words. The second exception involves JAM's totals for IN where both Portuguese and English contribute with an equal number of content words (6 each). Whereas, in the latter case, one might attribute this balanced content word contribution to the fact that JAM's overall totals for English and Portuguese tokens for IN are also more balanced (36 English tokens and 27 Portuguese tokens), in MEG's case her total number of tokens for each language for the TI group do not indicate such balance (41 English tokens as opposed to only 11 Portuguese tokens). Although an examination of the relevant utterances might shed light on these apparent discrepancies, such qualitative analyses are reserved for Chapter 6.

Looking now at the totals of English content words across the interaction types (column five), we find low numbers which vary between 0 and 7. In four of the groups (MT, CH, PG and FP), the total number of content words for either sibling does not go above 3. In the remaining three groups (TI, LA and IN), however, the range is between 3 and 7. Rather than this difference being a function of interaction type, if we again compare the total tokens for English in column three, what we find is that with smaller overall token numbers, content words are more likely to appear in the top twenty occurrences for English as a result of the proportionate decrease in other types of morphemes.

If we now look at the *types* of content morphemes which make up the totals for both JAM and MEG in each language, we are provided with a further indication of how the siblings make slightly different use of English and Portuguese in CS utterances. Looking at column 5 in the last row of the table we can see that both JAM and MEG have low totals of content words in English: JAM produces 20 different

types of content morphemes while MEG produces 19. This is out of a possible 140 (top 20 occurrences X 7 different interaction types) and such low proportions reflect the contribution of the Matrix Language in terms of content morphemes. However, despite this apparent similarity between the siblings, if we look at the distribution of these content morphemes in terms of nouns, verbs and adjectives we find a difference (see column 4, final row). Out of the total of 20 content morphemes for JAM, 10 are nouns, 8 are verbs and 2 are adjectives. Of MEG's total (19), only 5 are nouns, 13 are verbs and 1 is an adjective. If we consider that the most typical contribution from the Embedded Language are nouns, followed by verbs and then adjectives, we might expect the opposite order to be found in the Matrix Language i.e. nouns to be comparatively less frequent than verbs. This is true of MEG but not of JAM and could again be further evidence of the 'classicness' of MEG's CS when compared to JAM.

Regarding the Portuguese totals for content words (column seven in the final row), we find 49 for JAM and 84 for MEG (out of a potential 140). The occurrence of these high proportions of content words in Portuguese (when compared to English) clearly confirms the latter's role as the Embedded Language in both siblings' code-switching. And again we could claim that MEG's use of the EL appears to be more classic than JAM's: whereas for JAM 35% of the top 20 occurences are content words, for MEG this percentage reaches 60%. Furthermore, in terms of content types (see column six in the final row), whereas for JAM there is a roughly equal split between nouns (24) and verbs (21) with very few adjectives (4), for MEG the overwhelming majority are nouns (51), verbs and adjectives accounting for comparatively less of the occurrences (22 and 11 respectively). All of this evidence points to slight differences between JAM and MEG in terms of the nature (i.e. morpheme type) of the contribution of the Embedded Language to their CS utterances, differences which will be examined in their linguistic context in Chapter 6.

In the discussion of the word lists shown in Tables 11 (MT) and 12 (TI), it was pointed out that pairs of translation equivalents could be found for both siblings in their top 20 occurrences. This is an interesting finding if we consider that in classic CS each contributing language has a specific role to play in terms of the types of morphemes it contributes to CS utterances. If a particular word and its translation equivalent (such as and/e) appear within the top 20 most frequently occurring words this implies a certain symmetry and not asymmetry!

An examination of the word lists for the other interaction types revealed further pairs of translation equivalents in the top twenty most frequently occurring words for JAM and MEG. In total there were 18 different pairs found in JAM's word lists and 10 different pairs found in MEG's. The following eight pairs occurred in both of the siblings' lists: and/e; the/a,o; I/eu; no, don't, isn't/não; is, yes/é; to/para; that, which/que and look/olha. Although the specific frequencies of each of these equivalents and the remaining 12 pairs (10 for JAM and 2 for MEG) will not be discussed here, two general observations can be made. Firstly, the data reveals that JAM uses more translation equivalents than MEG, both in terms of types and frequencies. Secondly, despite this difference, for both siblings there is evidence of an asymmetry at work within the use of pairs of equivalent morphemes in their CS utterances: when code-switching with MOT, the English equivalent is more frequent and when code-switching with PAI the Portuguese equivalent is more frequent. Although such a finding appears to reflect what has consistently been found in previous analyses (that the ML/EL asymmetry is a function of addressee), one must still recall that the very presence of translation equivalents in the top twenty frequency lists is unexpected.

This section set out to examine the word lists of the different interaction types in the light of the 4-M Model. Through a comparison of the frequency of content words it has been possible to gather more evidence for the existence of the Matrix/Embedded Language asymmetry in the code-switching of JAM and MEG. This asymmetry was seen to be present in all the interaction types which strengthens the claim that the siblings' language use is more dependent on the interlocutor variable than on the nature of the interaction. Differences between JAM and MEG in the frequencies of content words occurring in the English and Portuguese lists served to place both children at slightly different points on the code-switching continuum, MEG relatively closer to the 'classic' end where the contribution of Embedded language items is more restricted in terms of morpheme types and their frequencies. The analysis of the translation equivalents proved to be interesting as it provided, at the same time, both evidence for and against the ML/EL asymmetry. While the comparative frequencies of each item in the pairs appear to reflect the ML/EL asymmetry, the very fact that these pairs can be found in the top 20 occurrences of the lists for each language goes against the idea that morphemes from the ML and EL are mutually exclusive. However, when we find that 8 of these pairs are shared by JAM and MEG it becomes evident that this is not a random phenomenon and that

such a pattern of occurrence needs further investigation in order to explain its existence. The utterance-level analyses in section 6.2 will shed light on this matter as we are able to see how these pairs actually occur in the CS utterances.

At this point it is important to remember that it is only due to the methodology of this study that such insights can be gleaned from the data. It would be implausible to *manually* carry out a token count of the different morpheme types used by the speakers. Without the ability to perform automatic analyses of the data the patterns discussed above would probably go undetected. Furthermore, it is also important to highlight that such specific analyses of the data in this study are only possible because of the system of language coding used in the LOBILL Corpus. Bilingual corpora without this type of coding would not provide such rich results for subsequent human interpretation and analysis.

5.1.4 Frequency word lists of the code-switching occurring in other speaker/interlocutor combinations

As already mentioned, due to lower CS token counts a word-level analysis per interaction type was not feasible for the speaker/interlocutor combinations other than those involving the siblings addressing their mother. However, by performing the FREQ analyses on all the files at once we are provided with frequency lists for other speaker/interlocutor combinations which are worthy of brief commentry. Although I will not present the actual lists here, they were scrutinized in the same way as the lists in the sections above: the number of content words in the top 20 occurrences were counted for each language and translation equivalents were noted. Firstly I will comment on the MOT's frequency lists (when addressing her children) and then I will present a brief interpretation of the lists resulting from the CS occurring between the siblings.

As seen in earlier frequency analyses (see Fig. 7) the overall CS totals for MOT are very low: when addressing JAM the total for English is 550 and for Portuguese it is 132; when addressing MEG the CS totals are 229 and 75 respectively. Despite this, however, a pattern can still be detected from the lists[81]. In terms of content morphemes what we find is strong asymmetry between the languages: when addressing JAM, within the top 20 occurrences there is only one

[81] The two basic command lines to output the word lists were the following: kwal @ +t%add +t*MOT +s"JAM" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5 and kwal @ +t%add +t*MOT +s"JAM" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@en>" +r5

content word (the verb *say*) in the English list and 10 content words in the Portuguese lists. Of these content words 8 are nouns and 2 are verbs. A similar pattern is found in the lists with MEG as addressee: 4 English content words (3 verbs and 1 adjective) and 11 Portuguese content words (9 nouns and 2 adjectives). This asymmetry supports earlier evidence that MOT uses English as the Matrix Language and Portuguese as the Embedded Language when addressing both her children. In terms of translation equivalents we find the pair no/não used 9/2 times when MOT code-switches with JAM and a different pair the/a, used 9/2 times when code-switching with MEG. Although the frequencies of the items are low in both cases there is a certain asymmetry to be seen here which mirrors that found in the frequency of the items of each of the pairs in the siblings' lists. How and when these translation equivalents are used by the mother will be seen in the utterance level analyses (7.1).

It is when we examine the word lists pertaining to the siblings' code-switching with each other that we find results which differ quite significantly from those discussed in all the sections above. Again, the total CS tokens are low in both cases: the English/Portuguese totals for JAM (when addressing his sister) are 123/164 and for MEG (when code-switching with her brother) they are 99/106. When we count up the content words in each list we find the following: in JAM's English list there are 5 content morphemes (5 verbs and 2 adjectives) and in his Portuguese list there are 6 content morphemes (1 noun, 3 verbs and 2 adjectives). The numbers for MEG are similar to JAM's: 6 English content words (4 nouns and 2 verbs) and 6 Portuguese content words (3 nouns, 2 verbs and 1 adjective). In terms of the morpheme distinction predicted by the 4-M model what we could say is that as there are equal numbers of content words being contributed by both languages there appears to be no asymmetry at work here. If neither English nor Portuguese is taking on the role of the Matrix or Embedded Language in the CS utterances the siblings address to each other, how exactly are these utterances structured? Will we find what Myers-Scotton terms 'composite code-switching' in their utterances with both languages contributing the grammatical framework? An examination of their CS utterances will reveal all.

In terms of the occurrence of translation equivalents what we find are 7 pairs for JAM and 5 pairs for MEG. The pairs and their frequencies can be seen in Table 14:

Table 14. Frequency of translation equivalents in the top 20 occurrences of the siblings' frequency lists when code-switching with each other.

Translation equivalents (English/Portuguese)	JAM	MEG
the/a, o	8/2, 7	13/3, -
and/e	2/2	4/2
I/eu	7/10	3/5
isn't, no/não	2, -/11	3/4
look/olha	3/5	-
is/é	3/3	-
wait/espera	3/3	-
then/aí	-	2/1

The first four pairs occur in the CS utterances of both JAM and MEG and these same pairs were also found in the CS tokens addressed to their mother. However, this is where the similarity ends. In Table 14 it is not possible to detect the same pattern of language bias that was found when looking at the frequency of each of the items in the previous tables. There is no pattern of asymmetry: at times the Portuguese equivalent is more frequent (see the pairs I/eu and isn't, no/não) and on one occasion it is the English item which is more frequent (see the/a,o for MEG). The fact that there are 3 pairs (in JAM's case) where the frequencies are equal and another 5 pairs where the difference in frequency is only 1 or 2 (most of MEG's), points to a general lack of asymmetry in terms of the contribution of both languages with regards to these translation equivalents.

Although JAM and MEG code-switch very little with each other, the evidence so far (the frequencies of content morphemes and translation equivalents) indicates that neither English nor Portuguese appears to be assuming a definite ML or EL role in their code-switched utterances with each other. This contrasts with the asymmetrical roles that the languages play when they code-switch with their mother and father. It will be interesting to see how these differences are manifested in their utterances (Chapters 6 and 7). Firstly, however, the focus of discussion will now turn to a frequency analysis of the codes that have been used to annotate the LOBILL Corpus. As will be seen, such an analysis has the potential to reveal patterns in the data which would not be apparent to the human eye without recourse to specialist software.

5.2 Frequency code lists

In Chapter 3 the types of codes used to annotate the LOBILL Corpus were described. While some of these codes belong to the standard CHAT transcription system, others were specifically designed by myself in order to maximise the potential for the analyses relating to my research questions. This section aims to show what a frequency analysis of these codes can reveal about different aspects of the code-switching practice of the bilingual speakers in the corpus.

The first code to come under analysis will be the specially designed postcode used to mark all code-switched utterances. Occurring on the main line after the speaker's utterance it takes the form of a plus sign followed by the letters e (for English) and p (for Portuguese), the order and quantity of the letters determined by how the two languages structure the CS utterance (see 3.2.1.2 for more details). The second code to be examined is the standard one used to mark retracings and reformulations. Although there are, in actual fact, two slightly different codes, [//] for retracings and [///] for reformulations, at this stage they will be analysed together and the results merged. The third code to be analysed is that used to mark errors: [*]. Placed immediately after the error on the main line, there is also the option to provide more information about the error on a dependent line (underneath the main line). However, in this section the frequency analyses will only look at the main line. The fourth code, [@tq], is another specially designed code, used to mark tag questions on the main line, whether they are in English or Portuguese. The fifth and final code to be investigated is one that in standard CHAT usage marks quoted speech. In the LOBILL Corpus the use of this code, ["] was extended to include other metalinguistic uses, as will be explained in section 5.2.5.2. Although a further code, the one used to mark the variants of the kinship forms relating to MOT and PAI (@m or @p), was also analysed, due to textual restraints the results of these analyses cannot be reported on in this dissertation[82].

The majority of the frequency analyses reported on in the following sub-sections were carried out by using KWAL to select the speaker's utterances, and then FREQ to find all occurrences of the specifed code. With regards to the first code under analysis (the CS postcode), the variable of interlocutor was also incorporated into every command line. This is because, as has become increasingly evident throughout the discussions in this chapter, certain code-switching patterns are so intrinsically related to the variable of addressee that a failure to incorporate this factor

[82] For access to this excluded material interested readers can contact the author.

could lead to erroneous interpretations. With regards to the remaining four codes, here the interest lies in making comparisons between the frequency of the codes in CS utterances as opposed to non-CS utterances. In contrast to the CS postcode, these four codes relate to phenomena that can occur in both monolingual and bilingual speech and as such I want to investigate whether any relationships can be found between their frequency of occurrence and the speaker's language mode. Therefore, the interlocutor variable was not automatically included in the analyses and was only incorporated if it were thought to be productive. In all cases, examples of the command lines used to perform the frequency analyses can be found in the footnotes.

In keeping with the way the results in the previous section were presented, for most of the codes analysed I will present the results of the siblings side by side, first those pertaining to JAM and then to MEG. Where the interlocutor variable was incorporated into the analyses, the relevant speaker/interlocutor combination will be specified. Although in most cases there was no need to truncate the frequency lists, mention will be made of those where the top 20 most frequent items form the focus of the interpretation.

5.2.1 An analysis of the code-switching postcode

As mentioned above, in order to produce frequency code lists for each speaker/interlocutor combination (JAM/MOT, MEG/MOT, JAM/PAI, MEG/PAI, MEG/JAM and JAM/MEG) KWAL was used to select the relevant utterances and then FREQ was used to provide a frequency list of all the variants found in the CS postcodes[83]. For this particular analysis the use of the asterisk (*) in the command line (see +s<+ *>" in the footnote) was important since, acting as a wild card, it allowed FREQ to find all possible combinations of the letters e and p. In the following sections I will present the results of these eight analyses, beginning with the first two which relate to the frequency of the postcodes in the CS utterances addressed by the siblings to their mother.

5.2.1.1 Frequency lists of CS postcodes of the siblings when code-switching with their mother

[83] kwal @ +t%add +t*JAM +s"MOT" +u +d | freq +s"<+ *>" +o

The table below shows the frequency of the different variants of CS postcodes found in the code-switched speech of the siblings when interacting with their mother:

Table 15. Frequency of CS postcode variants occurring in JAM and MEG's utterances when addressing MOT.

JAM - MOT	MEG - MOT
211 [+ ep] 106 [+ epe] 87 [+ pe] 39 [+ epep] 16 [+ epepe] 14 [+ pep] 8 [+ epepep] 8 [+ pepe] 2 [+ emfp] 2 [+ epepepepe] 1 [+ ep] 1 [+ emf] 1 [+ emfe] 1 [+ epemfe] 1 [+ epepepe] 1 [+ epepepep] 1 [+ epepepepepep] 1 [+ mfe] 1 [+ mfemfe] 1 [+ mfepe] 1 [+ pemfe] 1 [+ pepepep] 1 [+ pepmfp] 1 [+ pses] 1 [+ se]	134 [+ ep] 64 [+ epe] 26 [+ pe] 22 [+ epep] 5 [+ epepe] 5 [+ pep] 4 [+ es] 2 [+ ese] 2 [+ pepe] 1 [+ emf] 1 [+ se]
Total types 25 Total tokens 508	Total types 11 Total tokens 266

There is a great deal that can be gleaned from this table. Firstly, if we look at the totals (see final row), it is evident that, in terms of numbers of utterances (total tokens), JAM code-switches significantly more than MEG when addressing MOT: 508 of his utterances as opposed to 266 of MEG's contain code-switched material. For JAM this number represents 10% of all utterances addressed to MOT (5,054) and for MEG this is only 4% of the total number of utterances she addresses to MOT

(6,589)[84]. In terms of types of CS utterances, we can also see a significant difference: JAM uses 25 variants as opposed to MEG's 11 variants. This could be interpreted as meaning that JAM's CS patterns are more varied than MEG's. If we now examine the actual variants of the CS postcode and their frequency what we find is that although the numbers for JAM are consistently higher than for MEG, the first six variants are exactly the same for both speakers. Topping both frequency lists is the variant [+ ep] which means that the utterance began in English and finished in Portuguese. Whether this switch to Portuguese involved a single word or more cannot be determined by the code as each single letter may represent more than one word. This variant is used 211 times by JAM and 134 times by MEG, accounting for 41% and 50% of the total number of CS utterances. The second most frequent variant for both siblings is [+ epe] which means that after switching to Portuguese JAM and MEG switched back to English. This variant accounts for 21% and 24% of JAM and MEG's totals. The third most common variant is [+ pe] which involves a switch from Portuguese into English. Occurring 87 times for JAM and 26 times for MEG, the percentages amount to 17% and 10% respectively. For the remaining three variants which occur in both lists the percentages are almost the same for the siblings: the variant [+ epep] accounts for 8% of the total for both JAM and MEG and the other two variants ([+ epepe] and [+ pep]) represent approximately 3% each of JAM's CS utterances and almost 2% each of MEG's CS utterances.

The very fact that the top six CS variants are exactly the same for JAM and MEG and that they occur with similar decreasing frequencies points to similarities in the way the siblings are using English and Portuguese to structure their CS utterances when addressing their mother. As seen above the most common variant is [+ ep] which tells us that over 40% of the time both siblings start their utterances in English and then switch to Portuguese. If we add to this percentage all the variants which indicate that an utterance was begun in English (all those which begin with the letter e in the lists) we see this percentage increase to 77% for JAM and 87% for MEG. Based on all the analyses thus far discussed in this chapter, there seems to be little doubt that the siblings use English as the Matrix Language when code-switching their mother. Could the high percentages for English initial CS utterances be interpreted as further evidence for this claim? It seems plausible that, for the most

[84] The following two command lines were used to output the total number of utterances addressed by JAM and MEG to MOT: freq @ +t%add +t*JAM +s"MOT" +u ; freq @ +t%add +t*MEG +s"MOT" +u

part, the language in which a speaker initiates an utterance would be a good indicator of the role of that language in a CS utterance i.e as the ML. Clearly there are exceptions (note the 17% and 10% of JAM and MEG's utterances which begin in Portuguese ([+ pe]). However, this relationship would be worthy of futher empirical study in other bilingual corpora where coding permits this type of analysis. In terms of the present study the only evidence which might support such a relationship is that provided by the frequency code lists which tell us the variants most used by the siblings when code-switching with their *father* with whom Portuguese plays the role of the Matrix Language. Before analysing these lists, however, there are still some observations to be made regarding the variants in the table above.

So far it has been possible to establish from the frequency code lists that although JAM code-switches more than MEG when addressing MOT, the patterns of code-switching (as shown by the variants) are remarkably similar. In fact, the first six variants, which are the same for both siblings, together account for over 90% of all the CS utterances: 93% in JAM's case and 96% for MEG. However, it is important not to ignore the variants which make up the remaining percentages. What is clear from the table is that JAM's CS utterances include variants where several switches take place. For MEG the greatest number of switches is in the variant [+ epepe] which involves 4 switches and occurs on 5 occasions. While this same variant occurs 16 times for JAM's, there are, in addition, 16 more utterances which involve more than 4 switches. One of these actually involves 11 switches back and forth (the variant [+ epepepepepep])! This particular utterance will merit special examination in section 6.1. Although the occurrence of these CS utterances involving more than 4 switches are not numerous it does seem to indicate that apart from code-switching more frequently than MEG in terms of numbers of utterances, JAM also appears to code-switch slightly more *within* some of the utterances. Overall MEG's approach to code-switching appears to be more measured. Whether this is the result of a conscious effort on the older sibling's part remains to be seen.

One final observation relates to the appearance of variants containing the letters *mf* and *s*. With regards to the former, which stands for '*mixed form*', what one would find in this type of CS utterance is a word which shares material from both languages, for example, a Portuguese verb with an English inflection, such as andas ('he walks'), where the English third person 's' has been added to the Portuguese third person 'anda'. From the frequency code lists we can see that these mixed forms

are rare in the data. However, while in MEG's list there is only one isolated case, in JAM's list there are 10 cases. It may be possible that there is a relationship between the amount and frequency of code-switching and the occurrence of mixed forms but this cannot be tested in this corpus. In any case, the motivation behind the use of mixed forms is unclear: they may be the result of unconscious 'mixing' of morphemes or the result of a conscious decision to play with the languages involved. By looking at the wider linguistic context surrounding the use of mixed forms in the LOBILL Corpus it may be possible to ascertain their motivation. With regards to the occurrence of the letter s, which stands for 'Spanish', in JAM's CS utterances there are two cases ([+ pses] and [+se]) and in MEG's there are 7 cases ([+ es] x 4, [+ ese] x 2, and [+ se]). The use of Spanish by the siblings is somewhat surprising as neither was learning Spanish nor had any contact with Spanish speakers. However, all will be revealed when these utterances are examined in section 6.5.

If when code-switching with their mother, JAM and MEG use English to initiate their utterances 77% and 87% of the time, would we find similar percentages with regards to their use of Portuguese-initiated utterances when addressing their father? By specifying PAI as the addressee in the command line (see footnote 77 for the original command line) it is possible to answer this question, the results of such analyses being the subject of the next section.

5.2.1.2 Frequency lists of CS postcodes of the siblings when code-switching with their father

If we begin by looking at those postcodes which begin with the letter p, what we find is that out of JAM's total number of CS utterances (101), 82 begin with Portuguese. For MEG, 87 out of the 102 CS utterances begin with Portuguese.

Table 16. Frequency of CS postcode variants occurring in JAM and MEG's utterances when addressing PAI.

JAM - PAI	MEG - PAI
39 [+ pe]	36 [+ pe]
33 [+ pep]	34 [+ pep]
9 [+ ep]	13 [+ ep]
4 [+ pepe]	9 [+ pepe]
3 [+ es]	6 [+ pepep]
2 [+ epep]	1 [+ epe]
2 [+ ese]	1 [+ pepepe]

2 [+ pepep] 1 [+ epe] 1 [+ epepe] 1 [+ pepepe] 1 [+ pepepep] 1 [+ pepmfp] 1 [+ pmf] 1 [+ se]	1 [+ pepepep] 1 [+ sese]
15 Total types 101 Total tokens	9 Total types 102 Total tokens

Converted into percentages, these Portuguese-initiated utterances account for 81% and 84% respectively. So we do indeed find very similar percentages to those that were observed for when the addressee was the mother. It appears reasonable to posit, therefore, that there might be a relationship between the Matrix/Embedded Language asymmetry and the proportion of CS utterances beginning with either language: a high percentage would be indicative of the Matrix Language while a relatively low percentage would indicate the Embedded Language. Although such a relationship may appear obvious to other researchers investigating code-switching, by providing empirical evidence it is possible to lend valuable support to such a hypothesis and remove it from the realms of intuition.

It is not only the percentages of Portuguese-initiated CS utterances which are very similar for the siblings. The first four CS postcode variants are exactly the same for JAM and MEG and together account for 84% and 90% of all the CS utterances. In contrast to the results shown in Table 15, which revealed JAM as a more prolific code-switcher than his sister when addressing MOT, the table above appears to show more equal usage of code-switching when addressing PAI: the numbers of CS utterances are almost the same (101 for JAM as opposed to 102 for MEG). However, when these totals are seen as a proportion of the overall number of utterances addressed by the siblings to PAI, again we see that JAM does actually engage in more code-switching with his father when compared to his sister: 101 CS utterances account for 13% of JAM's total number of utterances (754) while for MEG the percentage is 9% (102 out of 1157)[85]. In methodological terms, it is important to highlight here that this facility to be able to calculate these proportions is only possible due to the coding in the corpus and the appropriate use of FREQ to perform

[85] These totals were provided in the output for the following two analyses: freq @ +t%add +t*JAM +s"PAI" +u and freq @ +t%add +t*MEG +s"PAI" +u

the relevant searches. Without such a facility, the interpretation of these (and other) lists may have proved erroneous.

Returning to Table 16, another observation to be made is that, as was the case for the CS utterances addressed to MOT, here also JAM uses a wider variety of CS variants (15 compared to 9 for MEG), his use of Spanish (coded with s) and two mixed forms (mf) accounting for this difference. And in terms of numbers of switches, the lists tell us that when addressing their father both siblings do not go beyond 6 switches and that CS utterances involving 4 or more switches occur infrequently: 6 times for JAM and 8 for MEG.

Despite the differences in raw numbers which can be observed when comparing Tables 15 and 16, the discussion above has shown that the frequency of patterns of CS utterances of the siblings when addressing MOT and PAI are comparable: while in both cases JAM is seen to engage in more code-switching and uses a wider variety of CS variants than his sister, their use of English-initiated and Portuguese-initiated utterances is remarkably similar. The resulting percentages have shown that there appears to be a relationship between the language in which a CS utterance begins and the role it plays in said utterance, namely that of the Matrix Language. If this holds true for the CS utterances addressed by the siblings to their parents what should we thus expect with regards to the CS utterances JAM and MEG direct at each other? It is to their results that the discussion will now turn.

5.2.1.3 Frequency lists of CS postcodes of the siblings when code-switching with each other

Up until now, previous analyses have been unable to identity the existence of a ML/EL asymmetry at work in their utterances. Let us now examine the lists below in Table 17 to see if this is reflected in how the siblings initiate their CS utterances: a more balanced proportion of English-initiated and Portuguese-initiated utterances might indicate this lack of asymmetry.

Table 17. Frequency of CS postcode variants occurring in JAM and MEG's utterances when addressing each other.

JAM - MEG	MEG - JAM
17 [+ ep] 16 [+ pe] 8 [+ pep]	19 [+ ep] 10 [+ pe] 2 [+ epep]

2 [+ epe] 2 [+ epep] 1 [+ epepep] 1 [+ pses] 1 [+ se]	2 [+ pep] 1 [+ pmf] 1 [+ sese]
8 Total types 48 Total tokens	6 Total types 35 Total tokens

Despite the low overall number of tokens (48 for JAM and 35 for MEG), there are still patterns to be seen in the lists. First of all we can see that the most frequent for both siblings is [+ ep], that is, an English-initiated utterance involving just one switch to Portuguese, with 17 occurrences for JAM and 19 for MEG. In second place, however, we find the reverse of this variant, [+ pe], occurring for JAM with almost the same frequency as the first variant (16 times) and less for MEG (10 times).

In JAM's case, the almost equal frequency of these top two variants do indeed appear to reflect a symmetry in the use of both languages in terms of initiating CS utterances when addressing MEG. If we also divide the other variants in JAM's list according to their first letter and add them to these first two variants we find that the overall proportions are almost equal: out of the total of 48 occurrences, 22 (46%) are English-initiated and 25 (52%) begin with Portuguese. This supports the evidence gathered thus far that has indicated a lack of ML/EL asymmetry in JAM's CS utterances directed to his sister.

Of the 35 CS utterances MEG addresses to her brother, 21 (60%) are English-initiated and 13 (37%) are Portuguese-initiated. Compared to JAM there is greater disparity in MEG's percentages which might indicate a slight bias towards English being chosen more frequently than Portuguese. However, there is no strong asymmetrical pattern in evidence, which, according to the prediction proposed, means that the roles of the Matrix and Embedded Languages are not clearly defined in MEG's CS utterances.

Of course, it is important to bare in mind that although the evidence from all the analyses so far points to the fact that there appears to be no strong contender for the role of the Matrix Language in either JAM or MEG's CS speech when interacting with each other, the data analysed was collected over time and in different situations. It may be, for example, that many of the English-initiated utterances occurred while in England or in the period immediately after returning from holiday: immersion in an

English-speaking context would understandably have a triggering effect on their 'choice' of language when interacting with each other. Even if this involved an 'involuntary' start in English followed by reformulation in Portuguese, the CS postcode would still state that it was English-initiated. Therefore, in terms of the relationship between the initial language of a CS utterance and the Matrix Language, at least for the siblings' interactions with each other, the data is problematic. Each utterance will need to be examined in the light of contextual factors such as location, period of time, interaction type and the presence of other bilingual or monolingual interlocutors.

Despite these reservations in interpreting the siblings' code lists in Table 17, I believe those pertaining to the interactions between the siblings and their parents do reveal patterns which lend support to the proposed relationships mentioned above. Such asymmetrical patterns of language use in CS utterances have been identified again and again in the different analyses carried out by FREQ, VOCD and WDLEN and it is expected that the qualitative examination of the utterances (in Chapter 6) will confirm these findings.

5.2.2 An analysis of the codes for retracings [//] and reformulations [///]

Throughout the LOBILL Corpus, standard CHAT symbols were used to code words which were followed by retracings or reformulations, whether these were carried out in the same language or not. Whereas a retracing involves the incorporation of a word or words used just before the retracing, a reformulation involves a complete change of material. The two example utterances shown below illustrate what this coding looks like in code-switched utterances.

(22)
*JAM: <(be)cause I[//]>[@en] <eu gosto muito desse>[@pt] . [+ ep]
Because I, I like this very much. F027: L426

In this first example when JAM switches from I to eu, the symbol used is [//] indicating a retracing. This is because although the word is in a different language, the meaning is the same. If he had simply repeated the English pronoun and not switched to Portuguese, this would have been a simple case of repetition and coded with [/]. In the example below, MEG begins in English but reformulates in order to (better)

express what she wishes to say. This reformulation includes two switches to Portuguese:

(23)
*MEG: <it just makes[///]>[@en] mata[@pt] your[@en] sede[@pt] . [+ epep]
It just makes, kills your thirst. F020: L191

From these two examples it is evident that an analysis of the retracings and reformulations that occur in the speech of the siblings has the potential to shed light on the relationship between this phenomena and code-switching. One question would be whether a speaker's code-switched utterances contain relatively more retracings and reformulations than their monolingual utterances. The first group of analyses performed in this section set out to investigate this matter by allowing comparisons to be made between the frequency of occurrence of the codes [//] and [///] in code-switched and non code-switched speech.

First of all, two separate analyses were performed on each of the four bilingual speakers. For both analyses KWAL was first used to select all the utterances pertaining to the speaker. For the first analysis these utterances were then sent to FREQ in order to output the total number of tokens coded by either [//] or [///][86]. In the second analysis FREQ was directed to focus only on those retracings and reformulations which occured in CS utterances[87]. The results of both analyses allowed the subsequent calculation of the percentage of retracings and reformulations which were found in only CS utterances[88]. This percentage was then cross-referenced to the results of previous frequency analyses which had revealed how much of each speaker's speech was made up of CS tokens (see data in Fig. 6).

5.2.2.1 Frequency results of the codes for retracings [//] and reformulations [///] in the mono and billingual utterances of the siblings and their parents

The chart in Fig. 23 shows the results of the two analyses for each bilingual speaker. Here we are not particularly concerned with the overall total of tokens for retracings and reformulations but rather the proportions of these tokens that occur in CS utterances. If we compare the four columns (that is, the four speakers) we can clearly

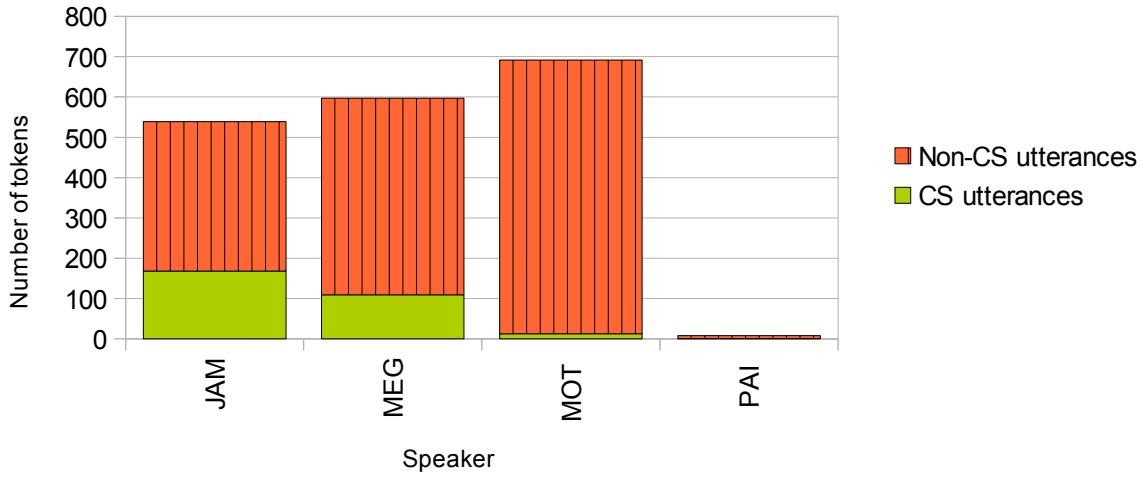
[86] kwal @ +t*JAM +u +d | freq +s"<//*>" +o

[87] kwal @ +t*JAM +u +d | freq +s"<//*>" +o +s"[+ *]"

[88] It was necessary to manually exclude some of the tokens from the overall token counts. This was due to FREQ's inclusion of certain codes in the word lists (such as [@en], [@pt], ["], [?], [@tq] and [: birthday]). The totals and percentages reported on in this section were all post-edited in this way.

see that it is in JAM's CS utterances that we find the highest proportion of retracings and reformulations: out of a total of 539 tokens, 168 occur in CS utterances, the latter accounting for 31%. For MEG the proportion is lower: 109 out of 597 converts to 18%. However, for MOT the percentage falls to 1.8% (13 out of 692 tokens). With regards to PAI, the numbers are too low (1 out of 8) to warrant further discussion.

Figure 23. Numbers of tokens of retracings and reformulations in non-CS utterances and CS utterances for the siblings and their parents.



The proportions shown in the chart indicate that JAM retraces and reformulates more than MEG when code-switching but that both siblings do so much more frequently than their mother. However, the data itself does not tell us whether more retracing and reformulating takes places in CS utterances or whether this phenomena is evenly distributed over non CS utterances and CS utterances (i.e independent of language mode). To make this comparison I looked back at the results of the analyses shown in Fig. 6. which revealed the percentage of tokens made up of code-switched material and included them in the table below (second column).

Table 18. Percentages of CS tokens and CS tokens which involve retracings and reformulations.

Speaker	Overall % of CS tokens	% of retracings and reformulations occurring in CS utterances
JAM	17%	31%
MEG	8%	18%

MOT	1.7%	1.8%

What a comparison of these percentages tells us is that retracings and reformulations are significantly more frequent in the siblings' CS utterances than in their monolingual utterances. If retracings and reformulations were proportionately distributed in the data the percentages in both columns would be the same or similar. However, for both JAM and MEG this is not the case: 31% of all of JAM's retracings and reformulations are found in only 17% of his overall token count (the CS tokens); and 18% of MEG's retracings and reformulations are found in 8% of her overall token count (the CS tokens). For MOT we see very similar percentages (1.7% and 1.8%) which indicate that despite overall higher token counts for retracings and reformulations (692) when compared to her children, these are evenly distributed across her utterances (CS and monolingual). That is, there is no evidence that she uses significantly more retracings and reformulations in her CS utterances than in her monolingual utterances.

What do these findings tell us? They indicate that for both JAM and MEG (but more so for JAM), retracings and reformulations are a significant feature of their CS utterances. It is likely that, as in the two example sentences, a retracing or reformulation may involve a switch into the other language. This could be the result of a repair strategy where JAM and MEG switch to the language favoured by their interlocutor after having 'involuntarily' begun in the other. It could be that the retracing or reformulation is the result of choosing a 'better' way of expressing a message or of the inability to continue the message in the same language. An examination of their CS utterances in section 6.2 will shed more light on the specific reasons behind the siblings' use of retracings and reformulations in code-switched utterances. As these utterance-level analyses will take into account the addressee variable it is pertinent to present here the results of a further set of frequency analyses which were carried out on four of the speaker- interlocutor combinations.

5.2.2.2 Frequency results of the codes for retracings [//] and reformulations [///] in the siblings' code-switches with their parents

By incorporating the addressee codes for the parents in the command lines, four more frequency analyses of the codes for retracings ([//]) and reformulations ([///])

were performed on the siblings' CS utterances (2 speakers x 2 interlocutors)[89]. The output of these frequency analyses can be seen below:

Table 19. Frequency of [//] and [///] codes in CS utterances addressed by the siblings to their parents

Speaker-Interlocutor	JAM-MOT	MEG-MOT	JAM-PAI	MEG-PAI
Retracings ([//]) Reformulations ([///])	54 14	30 7	22 12	15 11
Total frequency	68	37	34	26

The total frequency of the codes for JAM (68 and 34) reflect what was found above, in section 5.2.2.1: that he retraces and reformulates relatively more than his sister (37 and 26) when code-switching with his parents. What we now learn is that for all the above speaker-interlocutor combinations *retracing* occurs more frequently than complete reformulations, meaning that both siblings tend to favour repairs which incorporate material already uttered. This may, or may not, include the incorporation of translation equivalents: the totals in the table do not automatically mean that each of JAM's or MEG's codes necessarily represents a switch to another language, only that they occur in utterances where code-switching takes place. It is only be examining each of the codes identified above in their linguistic context that it will be possible to ascertain whether the retracings and reformulations are carried out in the same or different language, i.e whether they represent a switch point in the utterance. This investigation occurs in 6.2.

Before leaving this discussion on the analysis of retracings and reformulations in the LOBILL Corpus, in the section below I will briefly discuss a potential link between the results shown above and the MUL results that were discussed in section 4.3.2.

5.2.2.3 Cross-referencing of the retracings and reformulation code results with the MUL results

Through the WDLEN analyses it was possible to determine that for JAM, MEG and MOT the Mean Utterance Length values for CS utterances were significantly greater than that for monolingual utterances (refer to Table 10 for these means). For example, when addressing his mother, JAM's monolingual MUL (measured in

[89] kwal @ +t*JAM +t%add +s"MOT" +u +d | freq +s"[//*]" +s"[+ *]"

words) was 3.66 while his MUL for CS utterances was a higher 7.18. This pattern (a higher MUL for CS utterances) was observed for all the speaker-interlocutor combinations analysed. If, as evidenced by the analyses presented above, significantly more retracings and reformulations are seen to occur in CS utterances than in monolingual utterances, it is plausible to suggest that this might go some way to explaining the comparatively longer lengths of CS utterances when compared to the mean utterance lengths of monolingual utterances. By retracing and reformulating, a speaker is adding more words to their utterance and by doing this more frequently in CS utterances this would indeed result in a higher MUL for the latter.

I decided to carry out a simple cross-referencing of two different sets of data for speakers JAM, MEG and MOT. Whereas the MUL values mentioned above incorporated the variable of addressee (see 4.3.2.2), for the purposes of more general comparisons I decided to remove this variable, using WDLEN to output the MUL for monolingual utterances[90] and for CS utterances[91] for the three speakers, independent of addressee. In order to provide more samples for statistical testing I first divided the data set (119 files) chronologically into 12 groups of 10 files (the last group contained 9 files) and then repeated the two WDLEN analyses for each speaker per group. Rather than present all of the resulting 24 MUL values for each speaker, in Table 20 I show the overall mean utterance lengths of their monolingual (column 2) and CS utterances (column 3) and the overall difference between the means (column 4).

Table 20. Mean Utterance Length (MUL) results cross-referenced with retracings and reformulation results for JAM, MEG and MOT

Speaker	MUL of monolingual utterances	MUL of CS utterances	Difference between means	% retracings and reformulations occurring in CS utterances
JAM	3.71	7.87	4.16	31%
MEG	4.55	8.13	3.58	18%
MOT	4.68	5.77	1.09	1.8%

Both JAM and MEG average more than additional three words (4.16 and 3.58 respectively) for each CS utterance when compared with the average monolingual

[90]kwal @ +t*JAM +u +d | wdlen +r5 -s"@nonwords.cut" -s"[+ *]"

[91]kwal @ +t*JAM +u +d | wdlen +r5 -s"@nonwords.cut" +s"[+ *]"

utterance. For MOT the difference is just an additional one word (1.09). A paired t-test was carried out on the data for each speaker (12 monolingual MUL values versus 12 CS MUL values) and the results were found to be significant for JAM (t=-10.777, df=11, p<.001) and MEG (t=-4.912, df=11, p<.001) but not for MOT (t=-2.039, df=11, p=.066). Thus, for the siblings, the MUL values of their CS utterances are significantly higher than the MUL values of their monolingual utterances, confirming observations made earlier (see discussion of Table 10). If we now compare the percentage of retracings and reformulations occurring in the siblings' CS utterances (as opposed to in their monolingual utterances) with the difference in MUL means (column 4), we find evidence to support the hypothesis that there is a strong relationship between frequency of retracings and reformulations and longer CS utterance lengths. With 31% of JAM's retracings and reformulations occurring in his code-switched speech (17% of his data), he has the greatest difference in mean MULs (4.16). With a lower 18% of retracings and reformulations occurring in MEG's CS utterances (8% of her data), the mean difference in her MUL values is less (3.58) than her brother's. And with the occurrence of only 1.8% of MOT's retracings and reformulations in her CS utterances (1.7% of her data), there is only a very slight increase in difference between her monolingual and CS utterances (1.09 words). It is not possible to determine whether the differences between the siblings' (and mother's) CS MULs and monolingual MULs is the *exclusive* result of their retracing and reformulating in CS utterances. Nevertheless, there does indeed appear to be a relationship, one that has only come to light through the triangulation of the data resulting from separate analyses (with WDLEN and FREQ).

5.2.3 An analysis of the error code [*]

As mentioned in 3.2.3.3, throughout the LOBILL Corpus any error produced by a speaker is marked with the code [*]. As for the previous section on retracings and reformulations, here I am interested in investigating a possible relationship between the production of errors and code-switching, that is, whether more errors occur in code-switched utterances when compared to monolingual utterances. It is reasonable to argue that in the process of combining two languages in a single utterance, the potential for structural, lexical and phonological errors increases. This could be due to particular differences between the languages, such as differences in their syntax, their inflectional properties and their phonology. In addition cross-

linguistic influence could result in more errors related to the 'mis-use' of lexical items. Although frequency analyses will not reveal the reasons behind the occurrence of errors (see the utterance-level analyses in 6.3), they do allow us to make useful comparisons across the speakers and across language modes (monolingual or bilingual).

5.2.3.1 Frequency results of the error codes in the mono and bilingual utterances of the siblings and their parents

To begin with, two frequency analyses were performed on each bilingual speaker. In both cases KWAL was used to select the speaker's utterances and then FREQ was used to provide (i) the number of times the error code [*] occurred overall[92] and (ii) the number of times the error code occurred in only code-switched utterances[93]. As might be expected, due to the siblings' age difference, errors were seen to occur more frequently in JAM's discourse: there were 956 error codes for JAM and 371 for MEG. As one would also predict, the numbers of error codes for both parents were low: 19 for MOT and 4 for PAI. Bearing in mind that PAI's partcipation in terms of utterances is very limited, it does not make sense to use his results for comparative purposes and therefore the discussion in this section will focus on only the data of the siblings and their mother.

When we look at the frequency of error codes in the CS utterances for the three speakers we find the following: 180 for JAM, 48 for MEG and 0 for MOT. In terms of percentages of the overall occurrence of the error code, those occurring in CS utterances account for approximately 19% (JAM), 13% (MEG) and 0% (MOT). As these percentages relate to the frequency of the actual code and not the actual words coded as errors, two further analyses were carried out in order to output frequency lists for (iii) all the tokens coded as errors[94] and (iv) only CS tokens coded as errors[95]. Before examining the resulting word lists it is useful to compare the percentages arising from all four analyses, shown in the table below:

Table 21. Percentage of CS tokens, error codes and error tokens in CS utterances for JAM, MEG and MOT.

[92]kwal @ +t*JAM +u +d | freq +s"[*]" +o
[93]kwal @ +t*JAM +u +d | freq +s"[*]" +o +s"[+ *]"
[94]kwal @ +t*JAM +u +d | freq +s"<*>" +o
[95]kwal @ +t*JAM +u +d | freq +s"<*>" +o +s"[+ *]"

Speaker	Overall CS tokens	% error codes occurring in CS utterances	% tokens coded as errors occurring in CS utterances
JAM	17%	18.8%	19.5%
MEG	8%	12.9%	13%
MOT	1.7%	0%	0%

Looking first at MOT's percentages (final row), we already know that she code-switches very little (only 1.7% of the time) and now we know that of the errors she produces (19) none occur in her code-switched utterances (0%). When it comes to MEG (second row), the results tell us that almost 13% of all of her errors (371) can be found in her CS utterances (which account for 8% of her total token count). Although the difference between these two percentages is a little less than 5%, these results do indicate that MEG appears to produce more errors when she is in bilingual mode rather than when she is in monolingual mode. The fact that the percentages for MEG in column three and four are very similar (0.1% difference) means that for the overwhelming majority of each error code (48 in total) there is one single token (the error token total being 53). That is, MEG's errors mostly consist of single words.

The data for JAM reveals that he too produces more errors in bilingual mode than in monolingual mode: almost 19% of his error codes occur in his CS utterances which account for 17% of his total data. And the discrepancy between 18.8% (the percentage of error codes) and 19.5% (the percentage of error tokens) indicates that not all of his errors consist of single tokens. In fact, for his 180 error codes there are a total of 193 tokens, meaning that some errors must be referring to two or more words.

So far we have discovered that both children produce more errors in CS utterances than in monolingual utterances. We have also learnt that the proportion of errors involving more than one token is higher for JAM than for MEG. But what of the types of tokens marked as errors? By examining the frequency lists for each sibling it is possible to see the nature of the words which are coded as errors.

5.2.3.2 Types of tokens coded as errors in the siblings' code-switched utterances

With a total of 193 error tokens for JAM and 53 for MEG, the word lists will not be presented in their entirety. Rather, the top 20 occurrences will be shown and discussed below. From these lists it is not possible to know the exact nature of the

error as the tokens are removed from their linguistic context and here we do not have access to the dependent lines which describe each error in more detail (see 3.2.3.3). However, the lists do reveal differences between the siblings in terms of the types of tokens marked as errors (as well as their frequency of occurrence).

Table 22. Frequency word lists (top 20 occurrences) of tokens coded as errors in JAM and MEG's CS utterances

JAM	MEG
19 which	3 it
15 to	3 use
9 is	2 bombolê (hula hoop)
7 it	2 of
6 the	2 very
5 he	1 Robert
5 on	1 Sarah@pn
5 was	1 a
3 don't	1 amarelo (yellow)
3 got	1 at
3 he's	1 atolou (got stuck)
3 o (the, -m)	1 banco (bank, bench)
3 of	1 bit
3 want	1 caramba (gee)
2 anos (years)	1 carnivores (carnivores)
2 bateria (drum)	1 carpet
2 coleguinha (school peer)	1 christmas
2 de (of/from)	1 did
2 fall	1 drawed
2 his	1 estralando[96]
Types: 106 Tokens: 193	Types: 47 Tokens: 53

As can be seen in the table very frequent in JAM's list are the types which and to, occurring 19 and 15 times respectively. Neither of these appear in the top 20 of MEG's list, her most common error involving the words it and use, occurring only 3 times each. In fact there are only two types which appear in both lists, it and of, which indicates that the siblings' errors are, for the most part, different in nature. In terms of language, both Portuguese and English words can be found. However, of the top 20 occurrences, the majority are English with only 5 Portuguese types for JAM and 6 for MEG. In terms of types of morphemes, all of MEG's Portuguese words are content morphemes while two of JAM's Portuguese words are the masculine definite article o

[96] There is no translation for this word. It is likely that MEG wanted to say 'estalando' (cracking) and mistakenly added the 'r'.

and de. When we classify the English types we also find relatively more content words for MEG, JAM's list being dominated by grammatical morphemes. If we consider the frequency of the latter, one could say that the type of errors most frequently produced in CS utterances by JAM are grammatical in nature whereas the relatively smaller number errors produced by MEG appear to be more content based, that is, related to the lexical meaning of single items. The reasons behind this difference between the siblings are likely to be age-related: JAM's linguistic development in both languages is two and half years behind MEG's. With both grammatical systems less developed than MEG it does seem likely that JAM would be more susceptible to these types of errors when code-switching. It will be interesting to see whether this interpretation of the data will be supported by the examination of some of the errors in their linguistic context (see 6.3).

The analysis of the error code has proved to be very worthwhile. From the results of the four analyses it has been possible to make comparisons between the siblings in terms of the quantity of errors produced in CS utterances and the types of tokens involved. In JAM's case especially, the frequency lists have served to highlight certain words which are consistently being used erroneously, such as which and to. By examining these items in their linguistic context over time, it should be possible to see whether these types of errors disappear as JAM develops linguistically. As for the relationship between the occurrence of errors and code-switching, the evidence does suggest that when the siblings are in bilingual mode they are slightly more prone to the production of errors.

5.2.4 An analysis of the tag question code [@tq]

As explained in 3.2.2, the decision to code all the tag questions in the LOBILL Corpus was only taken after the transcription process was already underway. After noticing some 'deviant' uses of English tag questions by JAM in the first few transcripts, it was decided that this discourse feature was worthy of being coded for the purpose of future analyses. Of course, it is only possible to determine whether a tag question is being used appropriately or not by examining it in its utterance. Therefore, the investigation of tag questions in the LOBILL Corpus is more profitable

at an utterance-level analysis (see 6.4). However, a simple frequency analysis will provide some useful statistics about the use of tag questions by the speakers and pave the way for more indepth analyses later on.

As for some of the other frequency analyses in this section, FREQ was used to output the frequency of the tag question code per speaker in all utterances[97] and then in only CS utterances[98]. Frequency lists of the tokens coded as tag questions were then provided by two further analyses, one focussing on all utterances[99] and the other focussing on only CS utterances[100]. Before examining the resulting frequency word lists themselves (in section 5.2.4.2), I will first comment on what the quantitative results reveal about the use of tag questions by the four bilingual speakers.

5.2.4.1 Frequency results of the tag question codes and tokens in the mono and bilingual utterances of the siblings and their parents

The table below shows four sets of results per speaker: the overall frequency of the [@tq] code (that is, in all of the speaker's utterances), the frequency of the code in only CS utterances, the number of tokens (words) coded as tag questions in all utterances and the number of these tokens in only CS utterances. Percentages have been provided to aid interpretation of the results.

Table 23. Frequency results of the tag question code [@tq] for the siblings and their parents

Speaker	Overall frequency of [@tq] code	Frequency of [@tq] code in CS utterances (% of overall frequency of code)	Overall frequency of tokens coded with [@tq]	Frequency of tokens coded with [@tq] in CS utterances (% of overall frequency of tokens)
JAM	119	32 (26.8%)	164	44 (26.8%)
MEG	48	2 (4%)	64	2 (3%)
MOT	375	2 (0.5%)	650	3 (0.4%)
PAI	19	0 (0%)	21	0 (0%)

It is the results for JAM which stand out as being significantly different to the other three speakers. For MEG, MOT and PAI, the percentages in columns three and five tell us that the frequency of the tag question code and the tag question tokens

[97] freq @ +t*JAM +s"[@tq]" +u

[98] freq @ +t*JAM +s"[@tq]" +u +s"[+ *]"

[99] freq @ +t*JAM +s"<@tq>" +u +o

[100] freq @ +t*JAM +s"<@tq>" +u +o +s"[+ *]"

occurring in CS utterances are extremely low, or non-existent in the case of PAI. With only two tag questions being used by MEG in CS utterances, the resulting percentage of 4% falls below the 8% that one might expect if tag questions were distributed equally across the code-switched and non code-switched data. If we recall that MOT code-switches only 1.7% of the time, one would expect a similarly low percentage of tag questions in CS utterances. However, the percentage is even lower (0.5%) despite evidence that MOT uses lots of tag questions when in monolingual mode (375 occurrences). While the use of tag questions by MEG and MOT in CS utterances is virtually non-existent, for JAM we see a different pattern altogether. Out of 119 tag questions, 32 occur in CS utterances. The resulting percentage of 26.8% is almost 10% more than one would expect if we compare it to the percentage of CS tokens in his speech (17%).

Let us now turn to the actual frequency lists and examine the token types which are coded as tag questions in CS utterances. As PAI did not produce any tag questions in his CS utterances he has been automatically excluded from the table shown in the following section.

5.2.4.2 Types of tokens coded as tag questions in the code-switched utterances of the siblings and their mother

The contrast between the three speakers in terms of both total numbers of tokens and types coded with [@tq] is clearly evident in the table below.

Table 24. Frequency list of tag question tokens in JAM, MEG and MOT's CS utterances

JAM	MEG	MOT
12 isn't 12 yeah 11 it 2 né 1 he 1 is 1 não 1 tem 1 was 1 yes	2 yeah	1 is 1 it 1 yeah

1 é		
Types: 11 Tokens: 44	Types: 1 Tokens: 2	Types: 3 Tokens: 3

Despite such low token numbers for MEG and MOT, we do find one particular English token which occurs in all three of the lists. This is yeah?, an invariant tag question which JAM uses 12 times, MEG uses twice and MOT uses once. By subtracting these numbers from the totals for yeah found in the frequency lists for all utterances (not displayed here due to space restrictions) we find that this invariant tag also occurs frequently in monolingual English utterances: 16 times for JAM, twice for MEG and 42 times for MOT[101]. If we compare the frequency of yeah in terms of percentages we find that for JAM this tag occurs more frequently in CS utterances (27.2% of the time) than in non-CS utterances (13.3%). This increase in use might be due to the facility with which such an invariant tag can be used: unlike canonical tag questions which depend on previous material for their form, yeah is not bound by such restrictions. If in monolingual speech a speaker chooses this invariant over other forms because of its ease of use, it is not difficult to understand why it might be even more useful in code-switched discourse where previous material might be in a different language. The fact that both tag questions MEG uses in her CS utterances are yeah would also lend support to this notion.

Apart from yeah there are another two types which appear frequently in JAM's tag question frequency list for CS utterances: isn't and it. With 12 occurrences of isn't and 11 of it, it is likely that these are combined to form the tag question isn't it?. In non-CS utterances these two types are by far the most frequent for all three speakers: isn't appears 19 times for JAM, 8 times for MEG and 67 times for MOT; it occurs 26, 10 and 129 times respectively. It is perhaps not surprising, therefore, to find these two items at the top of JAM's list above. However, their occurence in JAM's CS list account for 27% (isn't) and 25% (it) of the total tokens while in his non-CS utterances these percentages are lower, 16% for isn't and 21% for it. This is potentially significant.

With regards to the use of Portuguese tag questions, we find two occurrences of the generic né (a contraction of não é which translates as isn't it) in JAM's CS

[101] As an alternative to subtracting from the totals, the following command line could be used to produce a frequency word list of all tokens occurring in non-CS utterances: freq @ +t*JAM +s"<@tq>" +u +o -s"[+ *]". Both methods were tested and produced the same results.

utterances. In monolingual Portuguese speech this particular tag performs like the English invariant yeah?, frequently being used instead of the canonical alternative. Overall MEG is the most prolific user of né in non-CS utterances (i.e. Portuguese utterances)[102]: it tops her frequency list with 21 occurrences, accounting for almost 34% of all her tag question tokens (62). For JAM the 16 occurrences of né account for 13% of all his tokens in non-CS utterances (120). If MEG shows such frequent usage of this tag question in her Portuguese utterances it is perhaps surprising that not a single occurrence can be found in her CS utterances whereas two can be found in her brother's. As it is, MEG's use of tag questions is restricted to two occurrences of the English yeah and nothing more.

The evidence provided by the frequency analyses of the code [@tq] point to clear differences in tag question usage between the speakers in both CS and non-CS utterances. As already seen, overall JAM uses tag questions more frequently than MEG (119 as opposed to 48) and a more than significant proportion of these occur in CS utterances. In terms of types of tag questions JAM shows a clear preference for two particular types when in bilingual mode, the invariant yeah? and isn't it?. It will be interesting to see in section 6.4 if and how these tag questions are actually combined with other language material in CS utterances.

5.2.5 An analysis of the metalinguistic code ["]

In the CHAT manual, transcribers are instructed to use the code ["] to mark single quote words in utterances. In the LOBILL Corpus, this code's use was extended to include the marking of any sort of metalinguistic reference, examples of which can be seen below:

(24)
*MEG: <how do you say>[@en] <batendo["]>[@pt] <in English>[@en] [+ epe]
How do you say 'hitting' in English? F024: L150

(25)
*MEG: +< <and why don't you call me>[@en] <filha["]>[@pt] ? [+ ep]
And why don't you call me 'daughter'? F108: L296

For utterances where more than single words were involved, the relevant material was enclosed in angled brackets and immediately followed by the code. In

[102] See the frequency list provided by the following command line: freq +t*MEG +s"<@tq>" +o -s"[+ *]".

interactions which involved the reading of stories (Literacy Activity interactions) it was often the case that entire utterances were coded with ["] as they were being 'quoted' from the story. The decision was taken to exclude all of these LA interactions from the analyses in this section as their inclusion would skew the results: hundreds of monolingual English utterances coded as quotes (i.e. the story being read by JAM or MEG) would not allow for a fair comparison of the use of quotes/metalinguistic comments in naturally occurring non code-switched speech as opposed to code-switched speech.

Before any analyses were carried out, therefore, all eleven LA files were removed at the file selection phase. Then KWAL and FREQ were used to output for each bilingual speaker (i) the overall frequency of the ["] code[103], (ii) the frequency of the code in CS utterances[104] (iii) a frequency list of the tokens coded by ["][105], and (iv) a frequency list of the tokens coded by ["] in CS utterances[106]. In addition, to allow for more reliable cross-referencing of the data, for each speaker the overall CS tokens percentages were recalculated[107] on the same files as above (i.e excluding the eleven Literacy Activity files)[108]. Before examining the word lists resulting from analysis (iv), the quantitative results of all five frequency analyses will be compared and discussed in the following section.

5.2.5.1 Frequency results of the metalinguistic codes and tokens in the mono and bilingual utterances of the siblings and their parents

When comparing the speakers' results, it is important to recall that whereas columns three and four in Table 25 refer to the frequency of the code itself, columns five and six refer to the actual tokens which have been coded by ["] whether that be a single word or a string of words occurring together[109].

---

[103] kwal @ +t*JAM +u +d | freq +s'["]'

[104] kwal @ +t*JAM +u +d | freq +s'["]' +s"[+ *]"

[105] kwal @ +t*JAM +u +d | freq +s'<">' +o

[106] kwal @ +t*JAM +u +d | freq +s'<">' +s"[+ *]" +o

[107] This was done by repeating the original two frequency analyses on each speaker (e.g. freq @ +t*JAM +u +o -s"@nonwords.cut" +r5 and freq @ +t*JAM +u +s"[+ *]" +o -s"@nonwords.cut" +r5) and then calculating the new percentages.

[108] Despite this concern for reliability, if we recall that *with* the eleven files the percentages of CS tokens were 17% (JAM), 8% (MEG), 1.7% (MOT) and 4.7% (PAI), their exclusion only resulted in a 1% increase for JAM and MEG. The difference for MOT and PAI was negligible. Clearly little code-switching took place in LA interactions.

[109] As for previous word lists, it was necessary to manually edit the lists by removing occurrences of the codes [@en] and [@pt]. Although the number of such occurrences was very low (ranging from only 1 to 5 per word list) such editing ensured more accurate totals and therefore more precise

Table 25. Frequency results of the metalinguistic code ["] for the siblings and their parents

Speaker	Overall CS tokens*	Overall frequency of ["] code	Frequency of ["] code in CS utterances	Overall frequency of tokens coded with ["]	Frequency of tokens coded with ["] in CS utterances
JAM	18%	290	39 (13.4%)	499	45 (9%)
MEG	9%	427	67 (15.6%)	971	140 (14.4%)
MOT	1.8%	466	57 (12.2%)	773	83 (10.7%)
PAI	4.7%	42	10 (23.8%)	94	17 (18%)

*These percentages exclude any CS data from LA files

If we look at the data for JAM, what we see is that while the metalinguistic code does occur in his utterances relatively frequently (290 times), the proportion of its occurrence in CS utterances (13.4%) is lower than 18%, the latter being what we would expect if the codes were proportionately distributed across non-CS and CS utterances. In terms of token frequency, the percentage is even lower: only 9% of the tokens coded as ["] occur in CS utterances. With 39 codes (column 4) marking 45 words (column 6), we learn that in most cases the codes were marking single words. The data is therefore telling us two things: that JAM engages in more metalinguistic language use while in monolingual rather than bilingual mode, and that when code-switching, his metalinguistic references mostly involve single words (1.2 words for every code). In non code-switched utterances he averages 1.8 words for every code.

Such finding contrasts with the evidence we see for his sister MEG. Overall metalinguistic code occurs more frequently in MEG's discourse than in JAM's (427 times as opposed to 290) marking a total of 971 tokens. Of this total, 67 codes (15.6%) and 140 tokens (14.4%) occur in her CS utterances. These proportions are higher than the 9% shown in column two and indicate that she makes relatively more use of this language device when code-switching compared to when she is speaking only English or Portuguese. Also in contrast to JAM, the ratio of words to codes is relatively higher: in CS mode MEG averages 2.08 words to every code and in non CS mode her average is 2.3. The evidence is telling us that MEG engages in more metalinguistic language use than her brother, especially when code-switching. It is

calculations.

possible that this increased use is related to her comparatively more developed language awareness and greater skills at manipulating her two languages. We will see in the utterance-level analyses how much of MEG's code-switching is actually purely due to the use of metalinguistic references (such as those shown in the examples at the beginning of this section) or other devices such as quoting another's speech.

It is perhaps the results for MOT which are most unexpected when we consider the pattern that has emerged so far in this section on code analyses. When analysing the codes for retracings, errors and tag questions in MOT's CS utterances, in each case the results showed very low frequencies of occurrence, accounting for between 0% (errors) and 2% (retracings) of all occurrences. These results correlated with the low proportion (1.8%) of MOT's tokens which are made up of CS tokens, meaning that these linguistic features (retracings and reformulations, errors and tag questions) are not used with more frequency in CS utterances when compared to monolingual utterances. However, from the evidence in the table above, this is clearly not the case for the metalinguistic code ["]. With a total of 466 occurrences of the ["] code marking 773 tokens, we actually see that a relatively high proportion of these occur in MOT's CS utterances: 57 codes account for 12.2% of all the occurrences (see column four) and 83 tokens accounts for 10.7% of all of the tokens (column six). Compared to 1.8%, which represents the contribution of CS tokens to MOT's overall token count, it is possible to say that metalinguistic language use appears to be a particularly significant feature of MOT's code-switching practice. The same could also be said of PAI if we look at his percentages: 23.8% of the ["] codes and 18% of the related tokens occur in his CS utterances. Even taking into account his CS tokens percentage of 4.7% (column two), which is higher than MOT, his percentages are still noticeably higher than expected, although we must bear in mind that his overall contribution in terms of tokens is very low.

The results above have shown that metalinguistic language use appears to be a particularly significant feature of the code-switched discourse of MEG and her parents. Although there is evidence that JAM makes use of this device, in contrast to his older sister and his parents, there is no indication that he makes more use of it when code-switching. In fact, the results point to lower usage of metalinguistic language in his CS utterances. Let us now see what an examination of the word lists will reveal about the nature of this metalinguistic usage.

5.2.5.2 Types of tokens coded with ["] in the code-switched utterances of the siblings and their mother

Table 26 shows the top twenty occurrences of words coded with ["] for JAM, MEG and their MOT. Again, due to the low number of tokens (17) for PAI, it was decided to leave the discussion of his metalinguistic usage to Chapter 7.

Table 26. Frequency word lists of the metalinguistic code ["] for JAM, MEG and MOT

JAM	MEG	MOT
4 Portuguese 3 português (Portuguese) 2 burnt 2 crazy 2 eggplant 2 inglês (English) 2 mar (sea) 2 skatista (skateboarder) 2 tram 1 Cathy@pn 1 James@pn 1 Thomas 1 acordou (woke up) 1 beep+beep 1 boom 1 buffer 1 burro* (donkey) 1 da (of the -f) 1 e (and) 1 gol (goal)	4 obrigado (thank you) 4 zero 3 a 3 and 3 burro* (donkey) 3 filha (daughter) 3 mantequilla* (butter) 3 na (in the -f) 3 say 2 animal 2 black 2 breathing+lungs 2 cabeça (head) 2 girando (turning round) 2 guinea+pigs 2 is 2 kittens 2 mucho* (a lot) 2 name 2 o (the -m) 2 pesado (heavy)	4 cinco (five) 3 amigo (friend) 2 Catarina@pn 2 Cathy@pn 2 Portuguese 2 Sara@pn 2 a 2 cadeira (chair) 2 dá (give) 2 escuro (dark) 2 mais (more) 2 mar (sea) 2 me 2 o (the -m) 2 português (Portuguese) 2 que (that) 2 tia (aunt) 2 vamos (let's go/we're going) 1 Telestunt 1 Visconde (Vicount)
Types: 33 Tokens: 45	Types: 103 Tokens: 140	Types: 62 Tokens: 83

*Spanish words

With regards to morpheme types, we find an overwhelming majority of lexically-laden items in all three lists. Out of the top twenty most frequent words there is only one grammatical token for JAM (na), three for MEG (o, a and na) and three for MOT (o, a and que)[110]. If we remove the proper names and Spanish words from the counts and

[110] Although the word 'a' can be found in both MEG and MOT's lists, without referring back to the transcripts it is not possible to ascertain whether this is the English indefinite article or the Portuguese feminine definite article. However, in both cases they have been classed as grammatical morphemes.

then examine the lists in terms of language, we find that for JAM and MEG the lists are quite balanced: there are 8 English words and 8 Portuguese words in JAM's list and for MEG the totals are 8 and 6 respectively[111]. This contrasts with MOT's list where 13 words are Portuguese and only 1 is English[112](the word Portuguese)! With MOT's list showing such a heavy bias towards Portuguese, I decided to examine her complete word list (all 62 types) and found that 44 of the types and 59 of the tokens (out of 83) were Portuguese. This contrasted quite dramatically with the total of 9 types and 10 tokens for English. An earlier frequency analysis[113] had established that the total token contribution of Portuguese to MOT's CS utterances was 228 words. Of this total we now know that 59 words are related to metalinguistic usage which actually accounts for over 25% of all MOT's Portuguese tokens. This is significant and reveals that one of the particularly important functions of Portuguese in MOT's code-switched utterances is to refer to language metalinguistically. Only by examining the utterances themselves (in 6.5.3) will it be possible to see exactly how MOT uses Portuguese to this end.

Further comments to be made about the lists include the occurrence of the words Portuguese and português. The former occurs 4 times in JAM's list and twice in MOT's list, and the latter occurs 3 times and 2 times respectively. There are also two occurrences of inglês in JAM's list, but none in MOT's or MEG's. Coded as metalinguistic usage, the utterances containing these words will be of special interest as they clearly make direct reference to the languages in some way. Another observation is about the appearance of three Spanish words in the top twenty of the siblings' lists: burro (JAM and MEG), mantequilla (MEG) and mucho (MEG). In section 5.2.1 when the CS postcodes were under analysis, the results revealed that a very small number of postcodes included the letter 's' (which stands for Spanish) as well as 'e' or/and 'p'. We now know three of the Spanish words that are represented by this letter. However, their use by the siblings is still rather baffling and will only become clear when we see these items in their linguistic context (see 6.5).

The analyses carried out in this section on the ["] code have been able to shed some light on the relationship between code-switching and metalinguistic

[111] As for the word 'a' (see footnote above), it would only be possible to determine whether the words 'zero' 'animal' and 'a' are English or Portuguese by examining the original transcripts where they would be coded accordingly.

[112] Here the words 'me' and 'a' cannot be classed as English or Portuguese as they could belong to either language (see previous footnote).

[113] freq @ +t*MOT +u +s"[+ *]" +o -s"@nonwords.cut" -s"<@en>"

language use in the speech of the main informants. Through the discussion of the results, we have learnt that whereas JAM appears to make less use of metalinguistic language in his code-switched utterances, the other three speakers show increased usage when in bilingual mode. In addition, whereas for JAM and MEG there appears to be no particular preference for either English or Portuguese when they make metalinguistic references in their CS utterances, MOT shows a very strong bias towards Portuguese - by cross-referencing frequency results we are able to discover that a quarter of all of MOT's CS Portuguese tokens are directly related to metalinguistic language use. Of all the functions code-switching can have, this clearly is one of the most prominent for MOT.

The discussions in this chapter have served a dual purpose. While the interpretations of the word lists have revealed more about the nature of the ML/EL asymmetry existing in the informants code-switching, the analyses of the codes have allowed for the proposal of relationships between code-switching and different types of linguistic phenonmena. Such relationships will be examined in more detail in the next chapter, the focus of which is an utterance-level analysis of the siblings' code-switching occurring in the LOBILL Corpus.

6. Utterance level analyses and results

From the discussion of the quantitative and word and code level results (Chapters 4 and 5) it has become clear that explanations for certain findings can only be sought at the level of the utterance and beyond (in the discourse). In this chapter, therefore, the data will be examined from a more qualitative perspective as we see how English and Portuguese actually interact with each other in the bilingual speakers' code-switched utterances. The choice of utterances analysed here has been guided by the code-level results from the previous chapter (5). As such, the current chapter will be structured in a similar same way, each sub-section presenting the analysis and interpretation of code-switched utterances pertaining to each type of coding: CS postcodes; retracing and reformulation codes; error codes; tag question codes; and metalinguistic codes. All example utterances are presented in full CHAT format with the addition of a gloss in English inserted under the addressee tier[114]. As in the previous two chapters, the command lines used to output the target utterances can be found in the footnotes, thus facilitating replication.

6.1 An utterance-level analysis of the CS postcodes

In section 5.2.1 several frequency analyses were carried out on the postcodes which mark each CS utterance in the LOBILL Corpus. As previous evidence had already shown that the siblings favoured English as the ML when code-switching with their mother and Portuguese when code-switching with their father, I was then able to use the frequency results of the siblings to support the proposal of the existence of a relationship between the language in which a CS utterance was initiated and its role as the Matrix Language in that particular utterance. However, the results also showed that for both JAM and MEG there were exceptions to this rule. That is, some utterances were found to begin in the 'other' language, the Embedded Language. For JAM, the percentage of CS utterances begun in the EL (113) represented 22% of the total number of CS utterances (508) when addressing MOT and 18% (18) of the total (101) when interacting with PAI. For MEG these percentages were 12% (33 utterances) and 14% (14 utterances) respectively (the total number of CS utterances

[114]The gloss does not form part of the original transcription and is used here to facilitate reader comprehension. Although the insertion of glosses throughout the LOBILL Corpus at the time of transcription would have been preferable, time contraints did not allow this to be carried out. However, this does not mean that such additions cannot be made in the future.

being 266 and 102). It is of interest to examine these exceptions to the 'expected' pattern at utterance level in order to understand why they occur.

To specify only those utterances which begin in the 'other' language KWAL was used to select the utterances occuring between a particular speaker and interlocutor and the string +**s"[+ p*]"** or +**s"[+ e*]"** was then added - the former would select all CS utterances beginning with Portuguese while the latter would search for those beginning with English. As a matter of course it was decided to send all the output to specifically named files in order to keep track of the results and be able to refer back to them whenever necessary. This was achieved by simply adding the string +**f(*speaker name*)postcodes** to each command line. For example, the results for JAM were saved in a file name with the extension **fJAMpostcodes**. By also including the string +**d1** in the command line, the file names and line numbers were automatically included in the output. A total of four analyses were carried out for each of the following speaker/interlocutor combinations: JAM/MOT, MEG/MOT, JAM/PAI and MEG/PAI[115]. As mentioned previously, due to the more limited amount of code-switching occurring between the siblings themselves and in interactions with MOT and PAI as speakers, the code-switched data for these speaker/interlocutor combinations will be examined in a separate chapter altogether (Chapter 7).

6.1.1 Portuguese-inititated CS utterances addressed by the siblings to MOT

The first two analyses provided two lists of Portuguese-initiated CS utterances addressed by the siblings to their mother: for JAM they amounted to 113 and for MEG the total was 33[116]. These utterances were then examined in detail in order to search for explanations as to why the siblings should initiate such utterances in the Embedded Language. Looking first at JAM's 113 utterances, there were only 18 where the use of the EL appeared to be involuntary or where the use of single words could be considered as not constituting a meaningful switch in the sense that a monolingual speaker of English would have no difficulty in comprehending such words. Three examples of such utterances are shown below:

---

[115] kwal @ +t*JAM +t%add +s"PAI" +u +d +s"[+ e*]" +d1 +fJAMpostcodes and kwal @ +t*JAM +t%add +s"MOT" +u +d +s"[+ p*]" +d1 +fJAMpostcodes

[116] To arrive at these totals one could either count up the utterances manually or perform the following type of frequency analysis: kwal @ +t*JAM +t%add +s"PAI" +u +d | freq +s"<+ p*>" +o. The use of the switch +s"<+ p*>" would provide a list of all those postcodes beginning with Portuguese (+ p*) in order of frequency, along with the total.

(26)
*JAM: é[@pt][///] yes[@en] . [+ pe]
%add: MOT
Yes, yes. F018: L238

(27)
*JAM: <o(lha)>[@pt] <there's>[@en] +/. [+ pe]
%add: MOT
Look, there's... F046: L604

(28)
*JAM: <não>[@pt], <she[/] she's going to>[@en] <trazer Samuel@pn[/] Samuel@pn>[@pt] . [+ pep]
%add: MOT
No, she's going to bring Samuel, Samuel. F086: L434

The first example shows a simple retracing of which there were three similar cases out of the 18[117]. The second example sees JAM beginning his utterance with 'o(lha)' of which there were four cases overall. As for JAM's use of 'não' at the beginning of a CS utterance, this occurred four times. The other seven utterances involved the use of the Portuguese kinship forms 'Mãe' and 'Mama' and the exclamatory markers 'ei' ('hey'), 'ai' ('ow'), 'oi oi' ('oy oy') and 'aí' ('so'), the latter occurring twice.

Whereas one could argue that the 18 cases mentioned above could be discounted from the total of JAM's Portuguese-initiated utterances (113) being as they do not show a 'purposeful' use of Portuguese, there are still the remaining 95 utterances to account for. That is, 95 out of 508 times JAM's use of Portuguese to initiate CS discourse with his mother cannot be explained by involuntary usage. Although it is important to remember that 77% of the time JAM does inititate his CS interactions with his mother in English and that this corresponds with his 'normal' use of English as the ML, some of the remaining 95 utterances do actually appear to reveal a more Matrix-like use of Portuguese. The three examples below show that Portuguese, and not English, is being used to frame the utterance:

(29)
*JAM: <a gente já está>[@pt] <in England, yeah[@tq]>[@en] ? [+ pe]
%add: MOT
We are already in England, yeah? F052: L37

(30)

[117] The relationship between retracings and the use of translation equivalents will be discussed in the next section.

*JAM: <ontem de noite eu sonhei>[@pt] <which[*]>[@en] <eu tirei>[@pt] <two[/] two teeth>[@en] . [+ pepe]
%add: MOT
Last night I dreamed which I had two teeth taken out. F086: L336

(31)
*JAM: <que tal eu brincar com isso e a gente brincar com isso aí>[@pt] <as well>[@en] ? [+ pe]
%add: MOT
What about me playing with this and we play with that there as well?
F053: L1254

In all three examples English is clearly taking on the role of the Embedded Language, contributing content morphemes ('England', 'two teeth') and an adverb ('as well') while Portuguese is providing most of the grammatical structure in addition to content meaning. There are cases, however, where the roles are not clearly defined, such as in the following examples:

(32)
*JAM: <eu caí>[@pt] <at Anderson@pn 's house>[@en] . [+ pe]
%add: MOT
I fell over at Anderson's house. F086: L358

(33)
*JAM: <é por isso>[@pt] <which[*] the sky is ve:ry>[@en] longe[@pt] ? [+ pep]
%add: MOT
Is that the reason which[*] the sky is very far (away)? F032: L1202

In the first example JAM uses Portuguese for the pronoun and verb but then a locative phrase which follows English, and not Portuguese, syntax. In the second example, both languages contribute to the grammatical structure of the CS utterance.

On examination of all the 95 Portuguese-initiated CS utterances we indeed find that in these utterances Portuguese appears to go beyond its EL role, at times sharing the role of the ML with English and at other times becoming the ML. In order to obtain a clearer picture of how both languages contribute to these particular CS utterances in terms of morpheme type, the output from the first analysis (see footnote 105) was passed through a frequency analysis which provided separate words lists for each language. This was achieved by simply adding the following strings to the original command line: | **freq +o -s"<@en>"** for the Portuguese list[118] and | **freq +o**

[118] kwal @ +t*JAM +t%add +s"MOT" +u +d | kwal +s"[+ p*]" +d | freq +o -s"<@en>"

-s"<@pt>" for the English list[119]. The results were indeed indicative of a more equal participation of both languages in terms of the grammatical structuring of these CS utterances: in addition to contributing similar numbers of words (390 for Portuguese and 318 for English), in the top 20 most frequent words, the lists shared thirteen grammatical morphemes which occurred with similar frequency. Although the most frequently occurring words for each list were different, 'eu' (I) topping the Portuguese list and 'the' topping the English list (both with 16 occurrences each), their equivalents in the other language were not far behind, 'I' occurring 8 times and 'o/a' (the) occurring 13/8 times. With the frequencies of the grammatical words decreasing in a similar fashion, this is further evidence to support the notion of a shared contribution to the structuring of these particular Portuguese-initiated CS utterances. Such equal contribution is not found in those CS utterances beginning with English. This was ascertained by repeating the above two analyses but replacing **[+ p*]** with **[+ e*]**[120]. Although 'the' still tops the English list and 'o' and 'a' appear in third and fourth place in the Portuguese list, the difference in frequency is significant - 148 occurrences of the English article as opposed to 16 and 14 for the Portuguese masculine and feminine articles. Such disparity highlights the predominance of English in providing the grammatical structure, that is, acting as the ML, in English-initiated CS utterances.

These frequency analyses have helped provide further evidence for the idea that clues for determining the ML of CS utterances can be found in the initial word(s) used by the speaker. In JAM's case, we have seen that when he addresses his mother with a Portuguese-initiated utterance, he is then more likely to make use of Portuguese to help frame his CS utterance, as seen in the examples above.

Returning to the analysis of the utterances themselves, it is important to search for possible motivations behind the initial use of Portuguese in those 95 CS utterances for which explanations have not yet been given. In order to do this it seemed that a better method would be to first examine MEG's smaller number of utterances (33) and then search for similarities and differences in the data for JAM.

Turning, then, to the results for MEG, previous analyses had shown that only 12% (33 out of 266) of MEG's CS utterances addressed to MOT were found to begin in Portuguese. What a qualitative analysis of the KWAL results in this section

[119] kwal @ +t*JAM +t%add +s"MOT" +u +d | kwal +s"[+ p*]" +d | freq +o -s"<@pt>"

[120] kwal @ +t*JAM +t%add +s"MOT" +u +d | kwal +s"[+ e*]" +d | freq +o -s"<@en>" and kwal @ +t*JAM +t %add +s"MOT" +u +d | kwal +s"[+ e*]" +d | freq +o -s"<@pt>"

revealed was that of these 33 utterances, six involved straightforward retracing of single words into English, indicating an involuntary start in Portuguese, three involved the use of 'o(lha)' (look) and a further two involved the kinship terms 'Mama' and 'Mamãe'. Out of the 33 Portuguese-initiated utterances, therefore, 11 can be accounted for in the same way as for JAM.

Of the remaining 22, however, a closer examination revealed that there were three cases where MEG's initial use of Portuguese was related to metalinguistic usage, as shown below:

(34)
*MEG: <liquidificador["]>[@pt] yeah[@en] . [+ pe]
'Blender' yeah? F070: L113

(35)
*MEG: <<sobre sobremesa>["]>[@pt], <it means>[@en] <<a gente está falando sobre mesa>["]>[@pt] . [+ pep]
%add: MOT
'About dessert', it means 'we are talking about table'. F078: L509

(36)
*MEG: +" <quem não comprar sapatilha ou pulseira colorida ou>[@pt] necklace[@en] +... [+ pe]
%add: MOT
'Whoever doesn't buy pumps or colourful bracelets or necklace...' F090: L181

In the first example, MEG is simply checking the Portuguese word for 'blender'. In the second example, she is talking about the word 'sobremesa' which as a compound means 'dessert' but separately means 'about' (sobre) and 'table' (mesa). In the third example she is directly quoting her (Brazilian) class teacher who had talked to them that day about their school dance[121]. Such metalinguistic usage was also found when JAM's utterances were re-examined: two cases involved quotations and another case involved JAM trying to explain what the word 'sagua' referred to, as shown below.

(37)
*JAM: +" há[@pt] <mucho burro>[@sp] on[@en] <dos três>[@sp] . [+ pses]
%add: MOT MEG
'There is a lot of donkey on two three'. F079: L632

[121] This contextual information was gleaned from consultation of the relevant part of the transcript which was achieved by using the following command line: kwal @ +s"sapatilha" -w5 +w5. The output showed 5 lines above and 5 lines below the specified key word (in this case 'sapatilha').

(38)
*JAM: +" socorro[@pt] Mum@m[@en] ! [+ pe]
%add: MOT
'Help Mum'! F096: L598

(39)
*JAM: <sagua["]>[@pt] <the name of that xxx>[@en] . [+ pe]
%add: MEG MOT
'Sagua' the name of that xxx. F108: L529

The first example is a direct quote from a character in the British sitcom Fawlty Towers (see section 6.5 for further explanation of this apparently nonsensical utterance) while the second example is a self quote occurring during the recounting of an incident. When examining the third example in its wider discourse context the referent of the word 'sagua' did not become clear but it became apparent that he was using it as a proper name. In both MEG and JAM's Portuguese-initiated CS utterances, therefore, we find three cases each where the use of Portuguese is directly related to metalinguistic usage. As will be seen in section 6.5, such usage accounts for a substantial amount of the code-switching occurring between particular speakers in the LOBILL Corpus.

Returning once again to MEG's CS utterances beginning with Portuguese, 14 out of the 33 have been accounted for above. Of the remaining 19 CS utterances, what came to light on closer examination was that MEG's use of Portuguese appeared to be related to sociocultural contextual factors. For instance, in the first example below, MEG's use of Portuguese is culturally bound: she plays hopscotch with her Brazilian friends and therefore perhaps does not know the English expression. In the second example she is referring to her father's favourite icecream and the Brazilian name would therefore be more used (and is also a lot easier to say) than the English equivalent.

(40)
*MEG: err[@en] <brincando de amarelinha>[@pt] <with <the rocks>[/] the rocks>[@en] . [+ pe]
%add: MOT
Err, playing hopscotch with the rocks, the rocks. F014: L51

(41)
*MEG: <flocos>[@pt], <where's our>[@en] <flocos>[@pt] ? [+ pep]
%add: MOT
Chocolate chip (icecream), where's our chocolate chip (icecream)? F076: L589

The fact that MEG refers to 'our' icecream is further indication of 'flocos' representing a shared referent for the family. The majority of these remaining 19 Portuguese-initiated CS utterances can be accounted for in this way, the use of Portuguese brought to the fore through its intrinsic links to everyday contexts such as school, television and meal times. When searching for such sociocultural usage in JAM's remaining 92 Portuguese-initiated CS utterances, only four cases were discovered. Three were school-related words; 'triángulo' ('triangle'), 'os Índios' ('the Indians') and 'Tia-Jeanne' *('*Teacher Jeanne'*);* and one was food-related; 'abacate' ('avocado'). This relative difference in occurences is quite surprising considering that JAM and MEG share the same sociocultural experiences. However, it is important to remember that we are looking at only Portuguese-initiated utterances and it may well be that socially and culturally embedded Portuguese words appear more frequently at other points in JAM's CS utterances.

Another examination of JAM's remaining 88 Portuguese-initiated utterances did reveal a category of words which appeared to occur significantly frequently in initial position: time related adverbs. With 14 cases in total, they were the following 8 Portuguese adverbs: 'depois' (then/afterwards/after) x 2, 'todo dia' (every day), 'quando' (when)[122] x 6, 'ontem' (yesterday), 'depois de amanhã' (the day after tomorrow), 'ontem de noite' (last night), 'amanhã′ (tomorrow) and 'agora' (now). Such usage was absent in MEG's Portuguese-initiated utterances, the only time related word being 'sábado' (Saturday), discussed below. This difference between the siblings was worthy of further investigation so I decided to perform further frequency analyses on the 8 Portuguese time-related adverbs mentioned above and on their English counterparts (in brackets). First I used FREQ and COMBO to output the occurrences of these words in all of JAM and MEG's utterances, COMBO being used to search for the multi-word expressions, such as 'todo dia'[123]. I then performed the same analyses on only the siblings' code-switched material: this necessitated the use of KWAL in order to select the CS utterances to be searched by FREQ and COMBO[124]. Although it would be interesting to make detailed comparisons across the data (i.e the frequency

[122] None of these cases were interrogatives.
[123] For example, freq @ +t*JAM +s"depois" +u and combo @ +t*JAM +s"todo^dia" +u
[124] For example, Kwal @ +t*JAM +u +d +s"[+ *]" | freq +s"depois" and kwal @ +t*JAM +u +d +s"[+ *]" | combo +s"todo^dia"

of each Portuguese and English time adverb occurring in JAM and MEG's output), textual restrictions only allow me to make the following observations.

Overall, MEG's frequency count of the time-related adverbs listed above came to 573, the English equivalents (430 occurrences) accounting for 75% of the total. For JAM this percentage was 66% (287 out of his total of 434). Such totals and percentages are a reflection of MEG's greater overall contribution to the LOBILL Corpus (43,428 tokens as opposed to 28,207 for JAM) and her greater use of English (over 10,000 more tokens) when compared to JAM. However, when we look at the results for only the CS utterances, as far as JAM is concerned the figures do not indicate such representativeness: the 121 occurrences found in his CS data represent 28% of the total, a percentage significantly higher than the 17% we might expect if they were evenly distributed across his non-CS and CS utterances. For MEG the number of occurrences of the time-related adverbs in her CS utterances (49) account for 9% of her total which is just one percent above her established benchmark of 8% (the percentage of her CS tokens). Thus, the 14 time-related adverbs found in JAM's Portuguese-initiated CS utterances addressed to his mother reflect the wider tendancy of JAM to use more of these types of expressions when in bilingual mode than when speaking monolingually. Interestingly, the frequency of two of the adverbs seem to be particularly high, with 'then' occuring 43 times and 'quando' occurring 30 times. If we add these together (73) they account for 60% of JAM's adverb total for CS utterances. Unfortunately, it is not possible to continue investigating these findings (by examining the pertinent utterances) as it is necessary to return to the focus of this section. However, it is clearly evident how such frequency analyses have the potential to allow for more indepth interpretations of the data at hand.

Returning to the results, of the remaining 74 Portuguese words with which JAM begins the CS utterances, the vast majority were grammatical in nature including, in order of frequency, auxilary verbs, personal pronouns, conjunctions, relative pronouns, prepositions and question words. As seen earlier on in this discussion, this highlights the role Portuguese is playing in these types of utterances; either that of the Matrix Language itself or of contributing to a shared 'composite' structure made up of both English and Portuguese.

In contrast, when we examine the structural make up of MEG's Portuguese-initiated CS utterances, it appears that, despite her use of Portuguese, she strives to

maintain English as the Matrix Language when addressing her mother as the following example shows:

(42)
*MEG: sábado[@pt] 's[@en] Sara@pn[@pt] <'s birthday at>[@en] Vovó@pn[@pt] 's[@en] . [+ pe]
%add: MOT
Saturday's Sara's birthday at Vovó's. (the Brazilian grandmother). F021: L224

Despite initiating with the Portuguese day of the week and referring to two Brazilian relatives, MEG still maintains the comparatively more complex syntax of English using the abbreviated 'is' and then the genitive ''s' twice![125] Such a felicitous combination of English and Portuguese in the face of two very different competing grammatical structures also serves to demonstrate MEG's ability to intertwine her two languages proficiently. Although this shows a clear preference to use English as the grammatical frame for her CS utterances, there are two exceptions, among the 19 utterances, where Portuguese takes on the role of the Matrix Language:

(43)
*MEG: <a vez da Mamãe@m>[@pt] <yeah[@tq]>[@en] ? [+ pe]
%add: JAM MOT
Mummy's turn, yeah? F063: L1158

(44)
*MEG: <eu odeio este cara aqui, parece um>[@pt] <burglar>[@en] . [+ pe]
%add: MEG JAM MOT
I hate this guy here, he looks like a burglar. F103: L581

The addressee lines shown above provide us with clues as to why Portuguese should feature so heavily in these two utterances: MOT is not the only intended addressee. And when examined within their wider context (i.e. by referring back to the original files) what we find is that both of these utterances were actually addressed to both MOT and JAM in separate interactions involving the playing of games (mini-snooker and the game 'Guess who?'). The first example appears to be more directed at JAM then MOT because MEG refers to 'Mummy' indirectly. If MOT is more of an indirect addressee in this case, that might provide an explanation for MEG's use of what is basically a monolingual Portuguese utterance with an English generic tag question. In fact, most of the exchanges between MEG and JAM in this

[125] The use of Portuguese syntax would have given rise to the following utterance: 'Sábado is the birthday of Sara at (the house of) Vovó'.

particular interaction (File 063) are in Portuguese, several involving the negotiation of turns. As for the second example, MEG is commenting in Portuguese on one of the 30 faces of the 'Guess who?' game and inserts the English word 'burglar'. While her opponent in the game is JAM, as her mother is helping JAM this comment is intended to be heard by both interlocutors. In both of these 'exceptions', then, it appears that the presence of more than one interlocutor is likely to have influenced MEG's use of Portuguese as the Matrix Language.

When we search amongst JAM's Portuguese-initiated CS utterances for those addressed to more than one interlocutor we find seven cases, six of which are addressed to both MOT and MEG and one of which is addressed to MOT and GRA (his English Grandmother). Of the former, five are framed by Portuguese with English contributing single words and the remaining utterance is the example shown above (14). It may well be that by including MEG as an addressee, JAM is prompted into using Portuguese as the Matrix Language. However, although this may be true for five of the Portuguese-initiated CS utterances, one must then ask why he also frequently uses Portuguese in such a way when his mother is the sole addressee. Furthermore, how do we explain the following CS utterance which JAM addresses to MOT and GRA:

(45)
*JAM: <para ser arma>[@pt] look[@en] . [+ pe]
%add: GRA MOT
To be a weapon, look. F031: L321

Despite being reminded three times during this particular conversation that his Grandma did not understand Portuguese, JAM continued to use Portuguese with her, sometimes as the ML (as in the example above) and sometimes sharing this role with English. An indepth analysis of the immediate discourse[126] revealed a possible motivation behind JAM's use of Portuguese: the subject matter of the conversation. The previous evening (in the dark) the family car had suffered a puncture and JAM is recounting this incident to his Grandma after which he begins making a car out of lego and putting weapons on it. Clearly excited by what had happened, it is likely that in his eagerness to recount the incident he draws on whatever language is more to

[126] Access to the relevant part in the transcript was achieved with the following command line: kwal @ +t*JAM +u +s"arma" -w3 +w3 + t%add. By adding the string '+t%add', the addressee tiers were also displayed in the output.

the fore, in this case, Portuguese. However, he is only able to do this because his mother is present as an active interlocutor (and interpreter) throughout the conversation – this frees him from the self-monitoring that would have slowed down his retelling of the incident. As we will see later, JAM's use of Portuguese here seems to be an isolated case - he does not code-switch with his Grandmother in any other files of the corpus.

Through the qualitative analysis and subsequent discussion of MEG's and JAM's Portuguese-initiated CS utterances addressed to MOT, it has been possible to uncover possible reasons as to why Portuguese features in these particular CS utterances. Although 'involuntary' uses and reformulations (indicating false starts) account for some of the utterances, other explanations lie in metalinguistic usage and the selection of Embedded Language items from the siblings' sociocultural context. However, as we have seen, in JAM's case there still remain several CS utterances (88) in which Portuguese features very heavily, the grammatical nature of its contribution frequently ousting English as the Matrix Language. The finding that JAM appears to favour Portuguese time-related adverbs, as opposed to English ones is an interesting one and worthy of further investigation in his remaining CS utterances. Again, through the qualitative analysis of the utterances it has been possible to further highlight the key role that the addressee and the presence of other interlocutors have in affecting the siblings' use of both languages in CS utterances. Without the specific addressee coding in the corpus, such insights would be lost.

6.1.2 English-inititated CS utterances addressed by the siblings to PAI

The third and fourth analyses[127] provided two lists of English-initiated CS utterances addressed by the siblings to their father: for JAM they amounted to 18 and for MEG the total was 14. These utterances were examined in the same way as those in the previous section in order to search for explanations as to why the siblings should initiate such utterances in the Embedded Language. Looking first at JAM's 18 utterances, there were 6 where the use of the EL could be classed as involuntary: 4 of these involved the retracing and subsequent translation of single words (and/e , but/mas, and yes/é x 2) and in the remaining 2 utterances JAM initiates with 'but' and 'no' before continuing in monolingual Portuguese. Such apparent false starts in the

[127] See footnote 115.

Embedded Language were also found in the JAM's *Portuguese*-initiated CS utterances addressed to MOT.

Metalinguistic usage can explain a further 5 of JAM's total (18): all of these involve direct quotes from the British sitcom Fawlty Towers where he code-switches from English to Spanish (see section 6.5 for more discussion). Portuguese does not feature at all in these 5 CS utterances and thus has no role to play, be it as the ML or the EL. There are now only 7 remaining English-initiated CS utterances for which involuntary or metalinguistic usage do not provide adequate explanations. One of these cannot be analysed as it contains the symbol xxx which means that a part of the utterance was not transcribed due to unintelligibility. In 3 of the remaining 6 utterances, English is firmly in the role of the Matrix Language, as can be seen below:

(46)
*JAM: <look at the light, it's very>[@en] preto[@pt] . [+ ep]
%add: MOT MEG PAI
Look at the light, it's very black. F039: L168

(47)
*JAM: <every day she goes when[///]>[@en] <quando atrapalho Mamãe@m>[@pt]
<every day she goes to my bed but sometimes she takes me and she leaves me there>[@en] . [+ epe]
%add: PAI MOT
Every day she goes when, when I bother Mummy every day she goes to my bed but sometimes she takes me and she leaves me there. F111: L400

(48)
*JAM: <and I'm>[@en] Sampaio[@pt] . [+ ep]
%add: MEG MOT PAI
And I'm Sampaio F111: L1077

Here, JAM's use of English, and not Portuguese, as the Matrix Language is not surprising given what we can see in the addressee tier: all three utterances are addressed to PAI *and* MOT. However, whereas in the first and third utterances one could potentially attribute JAM's use of English to his consideration of MOT as his primary interlocutor, in the second utterance this is clearly not the case. By making a reference to his mother (Mamãe@m) it is evident that his primary addressee is his father. Here a plausible explanation for his use of English as the Matrix Language lies in the influence of his linguistic environment and the change in language dominance resulting from his permanent move to England (see section 7.2 for more

evidence of this). Such influence can be seen in the remaining three utterances of the output for JAM's English-initial code-switched utterances with his father, discussed below.

Both of the following utterances occur in two separate telephone interactions while JAM is on holiday in England. In the first one (22 days after his arrival) JAM is telling his father about where his English cousins live: after initially beginning in English he then switches to Portuguese without difficulty. It is likely that his mention of Jake and Max triggered his initial use of English but Portuguese is clearly structuring the rest of the utterance as evidenced by JAM's use of the Portuguese genitive 'da'[128] when referring to his Grandmother's house:

(49)
*JAM: <when Jake@pn and Max@pn don't>[@en] <mora na casa da>[@pt]
Gran(d)ma@pn[@en] <<eles moram>[/] eles moram em outro canto>[@pt] . [+ ep]
%add: PAI
When Jake and Max don't live in the house of Grandma they live, they live in another place. F060: L204

(50)
*JAM: <(be)cause>[@en] <<a gen(te)>[///] eles>[@pt] <no[*] live in Gran(d)ma@pn 's house>[@en], não[@pt] . [+ epep]
%add: PAI
Because we, they no live in Grandma's house, no. F071: L303

Interestingly, a month later (almost 2 months after his arrival) English seems to have taken more of a hold on JAM's code-switching with his father, so much so that this time (in the second utterance) we see him using the English genitive ('Gran(d)ma's house') in an English phrase. Although he does use the Portuguese pronouns 'a gente' and 'eles' and inserts a token 'não' at the end, in this utterance it would be difficult to claim that Portuguese is maintaining its role as the ML. This is also true of the last of the CS utterances found in the output and shown below:

(51)
*JAM: <when we eat it we>[@en] acaba[@pt] <much more>[@en] <rápido>[@pt]
<than spaghetti>[@en] . [+ epepe]
%add: MOT PAI
When we eat it we finish much more quickly than spaghetti. F081: L53

Here, JAM inserts a Portuguese verb (acaba) and an adverb (rápido) in an otherwise English utterance, and does so successfully. The fact that his mother is also an

[128] In Portuguese, the genitive 'de' combines with the definite articles 'o' and 'a' to become either 'do' or 'da'.

addressee could explain why JAM is using English as a frame for this particular utterance.

From the analysis of the 18 English-initiated CS utterances JAM addresses to PAI it has been possible to establish reasons as to why JAM should begin a minority of CS utterances in the Embedded Language and not in the Matrix Language, as might be expected. False starts and metalinguistic usage accounted for 11 cases while a further 4 could be discounted due to the fact that PAI was not the sole addressee (this information is contained in the addressee tier). Removing the incomplete utterance, we were left with 2 CS utterances for which a purely surface-level linguistic analysis would not suffice. With access to contextual information it was then possible to correlate JAM's use of English in these two cases with the influence of his linguistic environment.

With regards to the proposal that the language of the initial word of a CS utterance is a good indicator of the Matrix Language of the same, a re-analysis of the above 18 English-initiated CS utterances actually reveals that in 11 of the cases this prediction is bourne out. This is illustrated in the five examples shown above where in all but one (example (49)) English takes on the role of the Matrix Language. It is no coincidence that in all of the remaining 7 CS utterances which appear to contradict the prediction, each initial English word (or phrase) could be considered an involuntary false start, triggered by the immediate linguistic environment. If we were to discount such CS utterances, the prediction described above would be further strengthened. One could even qualify this prediction further by stating that any exceptions to the expected pattern are likely to reveal involuntary code-switching. This interpretation is interesting if we consider the possibility of using a quantitative measure (i.e. the number of CS utterances which do not begin with the Matrix Language) to indicate the degree to which a speaker's code-switching is involuntary. A high proportion of CS utterances beginning with the Embedded Language would be indicative of more frequent involuntary code-switching while a low proportion might indicate a speaker's more proposital use of code-switching. It will be interesting to see whether the analysis of MEG's English-initiated CS utterances addressed to PAI lends further support to this notion.

An analysis of the KWAL output for MEG revealed that the reasons why she should initiate 14 of her CS utterances addressed to her father in English (typically the Embedded Language with this addressee) were similar to those found for JAM.

Her 'involuntary' use of 'yes', 'yeah' and 'no' at the beginning of otherwise monolingual Portuguese utterances, accounted for 3 of the utterances while a further 2 featured quotes (from Fawlty Towers). Four of the remaining 9 utterances included MOT as addressee and therefore one might well expect this to have influenced her use of English (whether in initial position or otherwise).

The influence of the sociolinguistic environment can be seen in 4 more of her English-initial CS utterances which were recorded while she was on holiday in England and talking to her father over the phone. Recounting her daily exploits, on one occasion she makes reference to the purchase of a set of dictionaries and uses the word 'dictionary' and then 'spelling dictionary' to initiate 2 of her utterances which she then completes in monolingual Portuguese (see F065: L234-310). Although she had already used the Portuguese equivalent ('dicionário') two utterances previously, her use of the English terms here could be considered almost metalinguistic, as if quoting the titles of a book. There is also little doubt that MEG is being influenced by her mother's use of the English terms as the latter responds to MEG's requests for further details about the purchase which she then relays to her father. The fact that MEG uses the English terms despite having already used the Portuguese equivalent does appear to indicate that she is aware that such usage does not hinder her father's understanding.

Such linguistic awareness is perfectly illustrated in the final case of MEG's use of the Embedded Language to initiate a CS utterance. I have chosen to show this particular utterance (underlined below) in its immediate linguistic context[129] as it is only by analysing it in this way that it is possible to explain MEG's use of the English word 'rock' in an otherwise monolingual Portuguese exchange with her father over the telephone. This excerpt shows MEG attempting to tell her father about a stick of rock which was bought during their day out at the beach. Both her mother and brother are listening in to her conversation, JAM waiting for his turn to talk to his father:

(52)
*MEG: <ah, e hoje na praia a gente comprou um>[@pt] <rock>[@en]. [+ pe]
%add: PAI
Ah, and today at the beach we bought a rock.
*MEG: <tu sabe, que tem o nome assim na frente que você fica chupando, chupando,

[129] Achieved with the following command line: kwal @ +u +s"rock" -w5 +w5 +t%add.

chupando e o nome nunca desaparece>[@pt].
%add: PAI
You know, that has the name on the front that you keep sucking, sucking, sucking and the name never disappears.
*PAI: www.
*MEG: <<rock, rock>["]>[@en] <<é um>[/] é um negócio assim duro>[@pt]. [+ ep]
%add: PAI
'Rock, rock' it's a it's a kind of hard thing.
*JAM: <<pedra, pedra[*]>["]>[@pt].
%add: MEG
'Rock, rock'.
%err: James translates word into Portuguese but it is the wrong meaning. Meggie doesn't use it because she is aware of the difference
*MEG: <um negócio assim duro>[@pt] +...
%add: PAI
A kind of hard thing.
*MEG: <é um negócio que a gente come>[@pt].
%add: PAI
It's a thing that we eat.
*PAI: www.
*JAM: <she's stupid>[@en].
%add: MOT
%com: James doesn't understand why Meggie won't translate 'rock' into Portuguese (pedra) to clarify what she's talking about F069: L137-158

Immediately after her first mention of 'rock', MEG tries to explain exactly what she is talking about by describing the unique feature of a stick of rock (that the name runs all the way through it). She is clearly aware of the potential ambiguity arising from the polysemy of the noun 'rock' and goes on to describe it as something hard and as something that we eat. Although JAM tries to intervene and 'help' his sister by supplying her with 'pedra', the Portuguese word for rock (as in stone), when this help is ignored he refers to her as stupid. JAM is evidently unaware that 'pedra' would not suffice in this situation. This excerpt is a perfect illustration of the differences between MEG and JAM in terms of the development of their linguistic awareness, an awareness that evidently has an effect on their code-switching practice, as will be seen in 6.5 when I compare the frequency with which both siblings code-switch for metalinguistic purposes.

When analysing MEG's 14 English-inititated CS utterances in terms of the Matrix Language prediction, what we find is that all of the 9 cases which appear to contradict the prediction (i.e that the initial word does not predict the ML of a given CS utterance) are either due to involuntary usage (3 cases), quoting (2 cases) or metalinguistic references (4 cases). Based on the data for JAM it was suggested

earlier that there might be a relationship between the number of exceptions to the prediction and the amount of involuntary code-switching taking place. While the data for MEG supports this notion to some degree (there being 3 cases of involuntary code-switching), one could not classify the functions of quoting and metalinguistic referencing as examples of involuntary code-switches. Indeed such propositional functions of code-switching are in direct contrast to involuntary uses such as false starts. Although it is necessary to therefore discard the proposed relationship mentioned above, the analysis of both JAM and MEG's data has been useful in being able to identify potential functions of the Embedded Language when found in inititial position in CS utterances - that of quoting and metalinguistic referencing.

The purpose of the analyses carried out in both this section and 6.1.1 was to search for explanations as to why JAM and MEG should inititate a minority of their CS utterances in the Embedded Language. Through the discussion of several examples it has been possible to highlight the motivations behind the occurrence of these exceptions to the code-switching patterns found in the majority of the data. While more conscious motivations included switching languages in order to quote somebody or refer to something metalinguistically, involuntary switches were seen to account for a significant number of the exceptions: of the total of 178 CS utterances analysed (131 for JAM and 47 for MEG), 38 involved involuntary switches (24 for JAM and 14 for MEG). Interestingly, in 14 of these cases the siblings were seen to immediately retrace and reformulate, providing translations of the original words. In the following section, a more detailed examination of all CS utterances involving retracing and reformulation will be carried out in order to shed more light on this particular feature of code-switching.

6.2 An utterance-level analysis of retracings and reformulations in code-switched speech

The frequency results discussed in section 5.2.2 revealed a relationship between code-switching and the use of retracings and reformulations in the utterances of JAM and MEG: the siblings appeared to retrace and reformulate more frequently when code-switching than when speaking monolingually. This was especially true of JAM whose frequency results showed significantly more use of this strategy when compared to his sister. More specific analyses revealed that when code-switching with their parents both siblings made more use of retracing than complete

reformulation. An additional finding which arose from triangulation with other data (the results from the WDLEN analyses) was that the relatively higher utterance length of the siblings' CS utterances, when compared to their monolingual utterances, could be attributed, in the most part, to this increase in use of retracing and reformulation. In this section the aim is to now examine the nature of these particular CS utterances and ascertain whether the retracings and reformulations are carried out in the same or different language, i.e whether they represent a switch point in the utterance. Through this examination I will also attempt to detect the reasons behind JAM and MEG's frequent use of this discourse strategy when code-switching.

KWAL was used to output the CS utterances the siblings addressed to their parents in which the retracing and reformulation codes occurred[130]. I will first discuss the siblings' CS utterances addressed to their mother before moving on to those addressed to their father.

6.2.1 Retracings and reformulations in the siblings' code-switches addressed to their mother

As the qualitative analysis of the output for MEG's code-switches with her mother proved to be more straightforward than that of her brother's I will begin with her results.

The data in Table 19 (see 5.2.2.2) showed us that when code-switching with her mother, MEG retraced and reformulated a total of 37 times. On examination of each code in the utterances themselves it was found that on 21 of these occasions MEG switched to a different language. In terms of the direction of the switch, most of them (13) occurred from Portuguese to English, the remainder (8) involving a switch in the opposite direction. The four examples below are typical of all of the 13 cases where MEG's retracing or reformulation involved a switch to English:

(53)
*MEG: eu[@pt][//] <I understand it <very little bit>[*]>[@en] . [+ pe]
%add: MOT
I, I understand it a very little bit. F026: L195

(54)
*MEG: <you said in>[@en] <inglês>[@pt][//] <in English, okay["]>[@en] . [+ epe]
%add: MOT

[130] kwal @ +t*JAM +t%add +s"MOT" +u +d | kwal +s"[//*]" +s"[+ *]" +d

You said in English, in English "Okay". F078: L614

(55)
*MEG: <oh I get confused with the>[@en] pa(lavras)[@pt][///] <<with the>[/] with the (.) words>[@en] . [+ epe]
%add: MOT
Oh I get confused with the words, with the, with the words. F026: L221

(56)
*MEG: <mmm <just the>[@en] <almo(ço)>[@pt]>[//] <just the eating part>[@en] . [+ epe]
%add: MOT
Mmm just the lunch, just the eating part. F049: L54

In the first three examples MEG replaces a Portuguese word with the English equivalent ('eu' becomes 'I', 'inglês' becomes 'English' and 'palavras' becomes 'words'), In the fourth example she replaces the Portuguese word for lunch ('almoço') with a paraphase ('just the eating part'). In these, and the remaining nine occurrences, it is evident that MEG is striving to maintain consistency in terms of the language she has chosen to use with her mother. To ensure this consistency, this at times means having recourse to the strategy of retracing and reformulation. However, there are times when maintaining language consistency is not easy or even desired, as indicated by the following three examples which are drawn from the 8 occurrences of retracings and reformulations which involve a switch into Portuguese:

(57)
*MEG: <<so they have to have a>[//] and they have to have a>[@en] chapeu[@pt] <and err>[@en] +... [+ epe]
%add: MOT
So they have to have a, and they have to have a hat and err. F010: L100

(58)
*MEG: Vitor@pn[@pt], <when I try to talk with him>[@en] <ele[//]>[@pt] he[@en][//] <ele me enche de chibata>[@pt] . [+ epep]
%add: MOT
Vitor, when I try to talk with him, he, he, he hits me. F026: L445

(59)
*MEG: <it just makes[///]>[@en] mata[@pt] your[@en] sede[@pt] . [+ epep]
%add: MOT JAM
It just makes, kills your thirst. F020: L191

In the first example MEG is talking about what her school peers have to wear for some up-coming celebrations. Her repetition of 'they have to have a' indicates that she

may have been searching for the English word 'hat', but, without being able to recall it easily, decides to use the Portuguese equivalent anyway. It is likely that this word is fresh in her mind from the morning's discussion at school (in Portuguese). In the second example, MEG is talking about a boy in her class at school who hits her when she tries to talk to him. Despite initially replacing the Portuguese 'ele' with 'he', MEG then uses 'ele' again and continues with a Portuguese phrase. This colloquial phrase translates literally as 'he fills me with hits' and carries more force than the word 'hit', perhaps explaining its use here. However, it is likely that the language of the school environment (Portuguese) again influences MEG's use of this expression: she may have already related the incident, in Portuguese, to her class peers and teachers. In the third example MEG appears to decide that a Portuguese idiom 'matar a sede' (kill one's thirst) better expresses what she wishes to say, although she does pay homage to English by skillfully inserting 'your' in the middle of the expression. The fact that JAM is also marked as an addressee might also indicate a further 'excuse' for MEG to use Portuguese.

The qualitative analysis of MEG's 21 cases of retracings and reformulations involving code-switching when addressing MOT was quite straightforward in the sense that there were no cases for which logical explanations could not be found. For her younger brother's data, however, the analysis proved to be more of a challenge, as will be seen in the following discussion.

In total, JAM retraced and reformulated on 68 occasions while code-switching with his mother. Of this total, 29 cases involved a switch in language, 11 from Portuguese to English and 18 from English to Portuguese. Note that this is the opposite of what was found for MEG - her retracings and reformulations were more frequent from Portuguese to English. Looking first at those utterances where JAM switches to English, what we find is that the retracings and reformulations appear to be straighforward cases of supplying the equivalent English word or phrase after having (involuntarily) begun in Portuguese:

(60)
*JAM: é[@pt][//] yes[@en] . [+ pe]
%add: MOT
Yes, yes. F018: L238

(61)
*JAM: <<agora>[@pt]>[//] now[@en] é[@pt] <Mister James@pn>[@en] . [+ pepe]

%add: MOT
Now, now it's Mister James. F092: L293

(62)
*JAM: <que mais eu vou>[@pt][//] <I'm going to give you>[@en] +/. [+ pe]
%add: MOT
What else am I going, I'm going to give you. F061: L346

(63)
*JAM: <porque eu tenho[///]>[@pt] <(be)cause I want to do>[@en] <mais um pouquinho assim>[@pt] . [+ pep]
%add: MOT
Because I have, because I want to do a little more like that. F034: L470

One might interpret these particular retracings and reformulations as a willingness to switch to his mother's mother tongue. However, subsequent switches into Portuguese (second and fourth examples) indicate that ease of expression might override this willingness at times. Indeed, it is when we look at the utterances where JAM switches from English into Portuguese (18 cases) that it becomes more apparent that JAM often finds it easier to express himself through Portuguese rather than English.

In the first two examples shown below, despite JAM's best intentions to make requests in English, after saying the initial modal words 'can' and 'could' he retraces and reformulates, switching totally into Portuguese to complete his request:

(64)
*JAM: can[@en][//] <posso te dizer>[@pt] . [+ ep]
%add: MOT
Can, can I tell you. F029: L889

(65)
*JAM: <Mãe@m>[@pt] cou(ld)[@en][///] <tu pode me dar um papel>[@pt][= sobbing] ? [+ pep]
%add: MOT
Mum, could, can you give me a piece of paper? F101: L354

The extra-linguistic information in the second example ([= sobbing]) tells us that JAM is obviously upset and the fact that he switches to Portuguese is an indication that he may find it easier to express himself in Portuguese, at least in emotionally charged situations such as this one. A further 8 cases are very similar to the above examples in the sense that the retracing or reformulation occurs very early on in the utterance

and that after the switch to Portuguese there is no return to English. That is, Portuguese takes over as the Matrix Language.

For JAM, difficulties (or unfamiliarity) with certain English structures are also cause for retracings or reformulations into Portuguese as can be seen in the following example.

(66)
*JAM: <no, I give him <to pres(ent)>[*][//]>[@en] de[@pt] present[@en] . [+ epe]
%add: MOT
No, I give him to present, as a present. F056: L505

In this utterance, JAM solves the problem of being unable to express 'as a present' in English by retracing and inserting the appropriate Portuguese preposition '**de**' which carries this meaning when used with the noun 'present'. However, it is rather puzzling to see that he actually uses the English word 'present' and does not use the complete Portuguese expression 'de presente', which might have been an easier and more harmonious choice. In the following example, JAM's difficulty in expressing himself leads to repetitions and retracings involving the same two words ('**to**' and '**the**'), thereby adding five extra words to his CS utterance.

(67)
*JAM: no[@en] melhor[@pt] <we go <to the>[/] <to[/] to the>[//] to>[@en] surfar[@pt]
<with the masks on>[@en] por(que)[@pt] +/. [+ epepep]
%add: MOT
No, (it's) better we go to the, to, to, to the, to surf with the masks on because...
F037: L43

This utterance is a perfect illustration of what the quantitative results had indicated - that the siblings tend to retrace and reformulate more in bilingual mode than in monolingual mode.

So far in this discussion the example utterances have illustrated that although there are occasions when JAM finds it easier to revert to Portuguese, either partially (by using a word or phrase) or completely (by completing the utterance in Portuguese), he does manifest a certain willingness to use English with his mother whenever he is able to. One might expect, therefore, that whenever JAM successfully expresses himself in English, he would have no need to call on his Portuguese. However, if this were so, how can we then explain the following three examples

where JAM retraces and provides his mother with the Portuguese equivalent of an appropriately used English word?

(68)
*JAM: <wait>[@en][//], <espera, espera>[@pt] . [+ ep]
%add: MEG MOT
Wait, wait, wait. F088: L312

(69)
*JAM: <<just gonna[: going to] eat beans>[//] just gonna[: going to] eat>[@en] feijão[@pt] <and this and[/] and>[@en] macarrão[@pt] <and then>[@en] +... [+ epepe]
%add: MOT
Just gonna eat beans, just gonna eat beans and this, and this spaghetti and then... F015: L516

(70)
*JAM: <can you imagine that if a head[///]>[@en] <cabeça>[@pt] <is so hard which[*] we if somebody[///] can we get an injection in our head>[@en] . [+ epe]
%add: MOT
Can you imagine that if a head, head, is so hard which we if somebody can we get an injection in our head. F077: L554

With regards to example (68) we find a feasible explanation for JAM's use of Portuguese in the addressee line: his sister MEG is also a target addressee of this utterance. However, there seems to be no pragmatic reason for JAM to provide the Portuguese equivalents found in the other two examples ('feijão' and 'cabeça') – there is no ambiguity in meaning which would warrant their use here. It may simply be that these particular words are so prevalent in JAM's linguistic repetoire (through everyday use) that they are uttered almost automatically, indicating again how more dominent a role Portuguese appears to play in his daily communication. Further evidence for this interpretation is provided when we examine the nature of the retracings and reformulations occurring in the CS utterances JAM addresses to his father, as discussed below.

6.2.2 Retracings and reformulations in the siblings' code-switches addressed to their father

In Table 19 (see section 5.2.2.2) we saw that there were 34 cases of retracings and reformulations in JAM's CS utterances addressed to PAI. On analysis of the utterances themselves[131] we learn that 13 of these cases involve a language switch,

[131] This output was provided by the following command line: kwal @ +t*JAM +t%add +s"PAI" +u +d | kwal +s"[+ **]" +s"[**//*]" +d

11 from English into Portuguese and only two from Portuguese to English. Of the former, seven involve the simple substitution of English conjunctions for Portuguese ones, as illustrated in the following three examples:

(71)
*JAM: <but>[@en][//] <mas, só que <o deles>[//] eles tem um debaixo e um de cima>[@pt] . [+ ep]
%add: PAI
But, but theirs, they have one downstairs and one upstairs. F064: L63

(72)
*JAM: <and>[@en][//] <e cavou muito que o[///] peguei o[///] a gente trouxe o>[@pt] +... [+ ep]
%add: PAI
And, and he dug a lot that the, I got the, we brought the... F069: L375

(73)
*JAM: não[@pt], <bec(ause)>[@en][//] <porque <não tem>[/] não tem bateria[*] não>[@pt] . [+ pep]
%add: PAI
No, because, because there isn't, there isn't a battery. F065: L55

A futher two cases involve JAM substituting 'yes' for the Portuguese equivalent 'é', as seen in example (74):

(74)
*JAM: <yes>[@en][//] é[@pt] . [+ ep]
%add: PAI
Yes, yes. F069: L366

There is even one case where JAM stops himself from saying 'Brazil', switching to the Portuguese pronunciation 'Brasil':

(75)
*JAM: <que no>[@pt] Braz(il)[@en][//] <Brasil tem um mais[/] mais>[@pt] fat[@en] . [+ pepe]
%add: PAI
That in Brazil, Brazil there is a more, more fat one. F109: 309

In 10 out of the 11 switches into Portuguese, the retracings appear to be straighforward substitutions indicating the desire, and ability, to supply the Portuguese equivalent after perhaps involuntarily having used the English term. One might ask why these English equivalents, and not the Portuguese ones, are selected

by JAM in the first place, given the fact that the evidence shows he is more at ease in Portuguese. However, it is worth remembering that most of the data for the combination JAM-PAI come from recordings of telephone calls between father and son while the latter was on holiday in England. Immersion in English would offer a plausible explanation as to why the English equivalent appears to be more readily accessible.

The effect of linguistic context can also be seen in the last of the English to Portuguese switches in JAM's data. He is telling his father over the phone about a recent thunderstorm and after using the word 'lightening' he then provides the Portuguese word 'trovão', which actually translates as 'thunder'.

(76)
*JAM: <não, mas[/] mas só que choveu e aí tinha um>[@pt] <lightening>[@en][//]
trovão[@pt] +... [+ pep]
%add: PAI
No, but, but it just rained and then there was a lightening, thunder... F064: 152

From this utterance alone it is not possible to confirm whether JAM believes he has provided the exact translation equivalent or whether this is additional content (i.e. lightening *and* thunder). This is where the advantages of corpus methodology come into play as we are able to quickly locate the utterance and examine it in its linguistic context. By simply typing **kwal @ +s"lightening"** into the CLAN command box, all the utterances containing this key word are displayed along with their file names and line numbers. The pertinent utterances were subsequently located in the files and clarification was thus made possible[132]. An analysis of the wider discourse revealed that although initally JAM appears to equate the two phenomena (that is, he *is* retracing in the example above), his subsequent replies to PAI's questions show that he quickly recalls the difference in meaning, telling his father that there was no lightening, just thunder. The fact that he continues to use the English term indicates that he does not know (or cannot recall) the Portuguese equivalent for 'lightening'. However, this does not present a problem for his father who clearly understands what JAM is telling him.

Of the two cases where JAM retraces or reformulates from Portuguese into English, one involves a culturally-bound referent, 'library'. While in England JAM and

[132] An alternative method is to use the 'w' switch which includes the surrounding dialogue in the output. The command kwal @ +s"lightening" -w5 +w5 +t%add, would output 5 lines above and 5 lines below the key word.

his mother would go to the local library in order to use the computer to send e-mails to PAI. After beginning to say 'computador', JAM then uses 'library', perhaps to clarify where they were using the computer.

(77)
*JAM: <eu disse para Mamãe@m no com(putador)[///]>[@pt] <library>[@en] <(es)tá fazendo assim o(lha) disse>[@pt] +"/. [+ pep]
%add: PAI
I said to Mummy in the computer, library, she's doing like this look I said. F064: L25

It is of no surprise that JAM does not use the Portuguese equivalent, 'biblioteca': these community facilities either do not exist or function in the same way in Brazil (at least in Fortaleza). It may even be that JAM was not familar with this word. In fact, a quick search for the words 'library' and 'biblioteca' in the whole corpus (achieved by the command lines **kwal +s"library"** and **kwal @ +s"biblioteca"**) revealed that JAM and MEG only ever used the word 'library' (2 occurrences for JAM and 4 for MEG).

The influence of the linguistic and cultural context may also provide an explanation for the second case where JAM retraces from Portuguese into English. Here JAM is recounting to PAI that his English aunt has a boyfriend.

(78)
*JAM: <ele>[@pt][///] <she has got a, erm boyfriend>[@en] . [+ pe]
%add: PAI
He, she has got , erm boyfriend. F074: L606

In this case, the contextual influence is strong enough to involve a complete take over of English as the Matrix Language. The fact that JAM initiates his utterance with an inappropriate Portuguese pronoun ('ele' instead of 'ela') further suggests that his Portuguese may be temporarily suffering through lack of use while in England. This interpretation is supported by the fact that this utterance occurs in a recording carried out at the end of the holiday, that is, two months after having arrived in England. In fact, on re-reading the transcript, one notes that JAM turns to his mother for help in recalling Portuguese words a total of five times while speaking to his father on the telephone. There is no evidence that his sister MEG requires the same assistance in the same situation, as will be seen in the following discussion of her results.

The frequency output for the combination MEG-PAI showed that MEG retraced and reformulated 26 times while engaged in code-switching with her father.

The output from KWAL then revealed that on only 11 of these occasions did she switch language: 7 times from English to Portuguese and 4 from Portuguese to English. Of the former, 6 involved the simple provision of the Portuguese equivalent after having already used the English word, as exemplfied in the these two examples:

(79)
*MEG: <tinha o[/] o>[@pt] <whale>[@en][//] <a baleia>[@pt] . [+ pep]
%add: PAI
There was the, the whale, the whale. F071: F231

(80)
*MEG: <<do outro>[@pt] <si(de)>[@en]>[//] <do outro lado>[@pt] . [+ pep]
%add: PAI
From the other side, the other side. F074: L189

In the remaining case MEG appears to belatedly recall the Portuguese word for dictionary and then retraces in order to correct her placement of the word 'French'. Athough she does not use the Portuguese equivalent of French, the structure of her utterance follows Portuguese requirements.

(81)
*MEG: <tinha[/] tinha[/] tinha>[@pt] <French>[@en][//] <um dicionário de>[@pt]
<French>[@en] . [+ pepe]
%add: PAI
There was, there was, there was French, a French dictionary. F065: L234

Of the four cases where MEG switches from Portuguese to English, the first one (shown below) involves the repetition of a complete phrase. MEG is talking to her father about a scene from the British sitcom Fawlty Towers where the Spanish waiter Manuel misunderstands his boss' instructions. In order to introduce a direct quote we see that she first uses a Portuguese reporting verb phrase and then the English equivalent immediately afterwards.

(82)
*MEG: <aí e[/] e na linguagem do Manuel@pn, ele pensa que é porcos e[/] <e ele
disse>[//]>[@pt] <and he said>[@en] +"/. [+ pe]
%add: PAI
Then, and, and in Manuel's language, he thinks that it is pigs and, and he said, and he said... F065: L344

On examination of the wider context of this utterance, we learn that she goes on to quote Manuel's question to his boss, "How did they get up there?"[133]). The fact that she is quoting English may have prompted MEG to have used the English reporting phrase. However, when introducing three more English quotes in the ensuing conversation, MEG only uses the Portuguese 'disse' ('he said'). Perhaps her use of the English equivalent in the example above served to make it clear to her father that she was quoting, and that once having established this, she was able to settle for just the Portuguese.

In the second example (shown below) MEG is recounting the siblings' visit to the beach. Although here we have a case of reformulation because there is a change in preposition from the Portuguese 'no' (in the) to 'to', no additional meaning appears to have been added by the insertion of 'to the sea'.

(83)
*MEG: <mas o James@pn, ele é maluco, se pulou no mar>[@pt][///], <to the sea>[@en] . [+ pe]
%add: PAI
But James, he is crazy, jumped in the sea, to the sea. F069: L87

Again KWAL was used to locate the file and line number of this particular utterance (**KWAL @ +s"sea"**) but the subsequent examination of the excerpt in the file did not shed any light on why MEG felt the need to add the English phrase. The use of the different preposition may indicate that this was a case of self-monitoring where MEG was not happy with the use of the Portuguese 'no' ('in the') to express what actually happened. Perhaps the preposition 'to' conveyed more appropriately the idea that JAM may have run into the sea and not jumped from a height, which is the image conjured up by the Portuguese phrase. This self-monitoring can be seen again in the third example where MEG is telling her father that while playing rounders (a British sport) with family relatives she was not caught out once.

(84)
*MEG: <e eu não fui pegada[*][/] pegada[*]>[@pt][\\] <caught out>[@en] <uma vez>[@pt] . [+ pep]
%add: PAI
I wasn't caught out, caught out, caught out once. F106: L293

[133] See File 065 lines 344-364 for the complete excerpt.

Marked as an error, MEG has used the incorrect past participle of the Portuguese verb 'pegar' (to get/to catch): despite being a regular AR verb in all other forms, the irregular past participle form which should be used here is 'pega'. After repeating the wrong form again (perhaps more for her own benefit as she notices that it sounds odd), MEG then uses the more specific expression 'caught out'. One has a sense that MEG is not satisfied with either the grammaticality or the lexical appropriacy of the Portuguese 'pegada' and rather than attempt to reformulate using Portuguese, chooses to use the expression which most accurately describes what occurred in this (new) sporting experience for her. There appears to be no problem with PAI's comprehension as in the following utterance MEG goes on to mention four more people that were 'caught out'. However, she manages to avoid the use of English again by just repeating the passive auxiliary 'foi' ('was') as in '[...] o Max foi, a Grandma foi [...]' (Max was, Grandma was (caught out)). It appears that the self-monitoring occurring here is more to satisfy MEG's own desire to express herself accurately and appropriately rather than out of pure consideration for her interlocutor. And the fact that she does not turn to her mother for any assistance indicates that she feels that she is able to gauge her interlocutor's comprehension perfectly well.

The fourth and final example of reformulation which involves a switch from Portuguese to English provides further evidence of MEG's concern for her accuracy of expression. MEG is telling her father about a dinosaur book her mother had bought for them that day and is describing one of the dinosaurs which she says looked like triceratops but did not have such a sharp horn.

(85)
*MEG: <era erm aquele que passa no filme <que tem aquela>[///] que parece>[@pt]
triceratops>[@en] <só que não é afiado a[/] a[///]>[@pt] <the horn>[@en] . [+ pepe]
%add: PAI
It was, erm, that one that was in the film that has that, that looks like triceratops, just that it is not sharp the, the, the horn. F065: L224

Her repetition of the Portuguese feminine article 'a' indicates that MEG is trying to think of the Portuguese word for 'horn' (chifre). Unable to do so she then supplies the English equivalent together with the definite article. The use of the English article 'the' after having already used the Portuguese article twice does not appear to reflect what normally occurs at such a switch point (between a definite article and noun). A

search for all occurrences of this particular switch point in MEG's CS utterances[134] showed that out of 27 cases there was not a single occurrence of MEG additionally inserting the English definite article between a Portuguese definite article and an English noun. Interestingly, and importantly for the analysis here, 25 of these cases involved the use of the Portuguese masculine definite article plus English noun (for example 'o deer', 'o bell', 'o dinner') which appears to be the default for this switch point for MEG, regardless of the gender of the Portuguese equivalent. This default is likely to arise out of the avoidance of possible grammatical ambiguity: the Portuguese feminine definite article 'a' has the same form (and often the same pronunciation) as the English indefinite article 'a'. In the two other cases where MEG uses the feminine definite article, they are both afixed to prepositions ('pra Beamish' meaning 'to Beamish') and 'na[//] no bricks' meaning 'on the bricks') thus not presenting any ambiguity.

Returning to the 'horn' example above it seems likely that, given the evidence, MEG has found herself in a singular position where she has no choice but to retrace and reformulate in order to be able to get her meaning across accurately: in this case the only way to do this is to use the English definite article in addition to the noun.

It is worth highlighting here that this interpretation was only made possible after investigating other occurrences of Portuguese definite articles followed by an English noun. This investigation was facilitated by the methodology used: without KWAL, the locating of these occurrences (via manual means) would have been laboriously slow.

Such an indepth investigation of this last example of MEG's was clearly necessary in order to arrive at a satisfactory interpretation of this rather singular case of retracing. In most cases, however, for both MEG and JAM, such detailed analyses were not needed as the explanations for their use of retracings and reformulations in their CS utterances with their parents were more straightfoward. Some of the examples shown revealed that the siblings had recourse to this strategy for similar purposes, one of these being to accomodate to their interlocutors' linguistic preferences, despite the influence of extra-linguistic factors (such as the environment and presence of other speakers) which may have initially triggered the 'other' language. MEG appears to be equally successful whether retracing or reformulating

[134] Achieved through the command lines **kwal @ +t*MEG +s"o" +s"[+ *]"** and **kwal @ +t*MEG +s"a" +s"[+ *]"**

into English or Portuguese, her examples showing evidence of proficient self-monitoring and effective self-repair as she strives for linguistic consistency with each interlocutor. JAM, on the other hand, clearly finds it more of a challenge to maintain this type of consistency, especially with his mother. Despite his best intentions, when retracing or reformulating with MOT, he is frequently seen (on 18 out of 29 occasions) to revert to Portuguese in order to be able to express himself. Such evidence seems to point to the fact that JAM is more proficient in Portuguese than in English and this in turn offers a possible explanation as to why he code-switches more than MEG when interacting with MOT.

In this section, the analysis of JAM and MEG's CS utterances involving retracings and reformulations has proved to be extremely enriching and is helping to build a personal profile of each sibling's code-switching behaviour. It has also served to provide an explanation for the rather unexpected presence of translation equivalents in the word frequency lists of the siblings, as revealed in 5.1. If we recall, the very fact that these pairs were found in the top 20 occurrences of the lists for each language went against the idea that morphemes from the ML and EL were mutually exclusive. From the examples examined in this section, it is now evident that most of the translation equivalents actually occur together, the result of involuntary usage immediately followed by the provision of the equivalent word in the other language. Such an insight may prove useful for other researchers investigating the ML/EL asymmetry in code-switched data, especially for those who may be basing their analyses, and subsequent interpretations, purely on word lists (i.e with no access to, or consideration of, the utterances themselves). Although the examination of word (and code) lists can be extremely fruitful (as shown in Chapter 5), it is through a more holistic, qualitative approach (as shown in this chapter) that we learn more about the different aspects of an individual's code-switching behaviour. In the following section, which looks at the siblings' errors in CS utterances, we will learn more about their ability to juggle their two languages.

6.3 An utterance-level analysis of the siblings' errors in code-switched speech

From the analyses performed in 5.2.3 we discovered that both JAM and MEG appeared to make more errors when in bilingual mode than when speaking monolingually. Another finding was that while MEG's errors were more lexically based, JAM's errors were much more frequently related to grammar. It was posited

that this was likely to be due to the fact that JAM's two grammatical systems were less developed than his sister's. In this section the aim is to examine all the error codes which occur in the siblings' CS utterances (180 for JAM and 48 for MEG) in their linguistic contexts and investigate this apparent relationship between the production of errors and code-switching. Due to the longitudinal nature of the data it seemed logical to analyse the utterances in chronological order: this would allow for the tracking of re-occuring errors and provide a more developmental perspective of the data. With this in mind, instead of initially selecting specific speaker-interlocutor CS utterances and analysing the output per interlocutor, I instructed KWAL to select all of each sibling's CS utterances which contained error codes irrespective of who they were addressed to - these utterances would automatically be listed in chronological order. Also included in the command line were two further strings, **+t%add** and **+t%err.** The former would output the addressee(s) of each output and the latter would ensure that any dependent lines coded as such would also appear in the output[135]. As explained in 3.2.3.3, the **%err** dependent line was used to add comments beneath utterances containing error codes and, where possible, the target form was also indicated. By enriching the transcription in this way, apart from facilitating my own analyses, other researchers would be able to make use of the data more effectively, particularly those whose knowledge of Portuguese is limited.

Returning to the command line, it is important to mention that the string **+d** was excluded. Normally used with KWAL to 'clean' the data of file names and line numbers in order for it to be piped to a second analysis, for this particular investigation this information was important. Therefore, this string was not used in any of the analyses carried out in this section.

From the difference in number of error codes occurring in JAM and MEG's code-switched utterances (180 compared to 48), as expected, the ouput for MEG was significantly less than that for JAM. However, even in her case, due to textual confines, only a reduced number of examples will be presented and discussed in this section. It is also important to bare in mind that the aim of this section is not to show detailed error analyses. Rather, I wish to highlight the effect that the use of two languages in a single utterance has on the type of errors produced by each sibling. In addition to the gloss, for each example presented I have placed the child's age in brackets as this will be important for comparisons across the siblings' data. The age

[135] kwal @ t*JAM +u +s"[*]" +s"[+ *]" +t%add +t%err

of any speaker can be calculated via the CLAN command DATES which uses two time values and computes the third[136]. The age of each child can also be found in the ID header at the beginning of each file.

6.3.1 Errors in MEG's code-switched utterances

Beginning with MEG's CS utterances, the first error identified by KWAL occurs in file number 007, meaning that in files 001 to 006 there are no errors in her CS utterances. In this first example below, MEG is talking to her mother about a picture she is painting and inserts the Portuguese word for 'brown' in an otherwise English utterance.

(86)
*MEG: <so I did[*] here (be)cause it looks like>[@en] marrom[@pt] . [+ ep]
%add: MOT (6;1.01)
%err: no pronoun
So I did here because it looks like brown. F007: L80

Her code-switch appears to be unrelated to her error which involves the non-use of the object pronoun 'it'. Whereas English grammar requires the use of the pronoun in this case, in Portuguese it is most common to only use the verb, the inclusion of the object 'o' being reserved for more formal speech and written Portuguese. It is likely that MEG's non-use of the object pronoun is simple transference from Portuguese.

In this second example we see that MEG's error occurs as a direct result of her code-switch. She is talking about what will happen when they try to remove a lamp post from the garden at the beach house and is concerned about the tree nearby.

(87)
*MEG: <it's gonna[: going to] fall on top of the>[@en] pobre[@pt] <of[*] the tree and the trunk>[@en] ! [+ epe]
%add: MOT (6;10.08)
%err: Portuguese transference
It's gonna fall on top of the poor of the tree and the trunk! F021:L331

Wanting to express 'the poor tree', a relatively straightfoward structure in English, she chooses to use the Portuguese word for 'poor' (perhaps thus intensifying the tree's predicament). However, it is clear that her use of 'pobre' has subsequently triggered

[136] The child's birthday and the date of each recording were inputted in the commands window in the following format: dates +b 07-OCT-1995 +d 08-NOV-2001. The output provided the age of the child.

the Portuguese structure 'pobre da arvore' which MEG then translates into English. This example shows transference from Portuguese but unlike in the first example (54), it is the code-switch which has caused this error.

The following three CS utterances support the finding from earlier frequency analyses that in the majority of cases MEG's errors are lexically based. All three utterances are grammatically well formed despite several switches between English and Portuguese which at times result in some retracings and reformulations. The errors, which involve two nouns and a verb are the direct result of the influence of the other language in terms of meaning. In the first of these three examples, MEG is talking about her school peer who had taken a large number of newspapers to school. The Portuguese word for newspaper is 'journal' (the plural being 'journais') and this is clearly foremost in MEG's mind as she simply uses the English pronunciation of the word and does not (bother to or manage to) recall the English equivalent of 'newspapers'.

(88)
*MEG: (be)cause[@en] Rafael@pn[@pt], <I think he bought ten or>[@en] <vi(nte)[/] vinte>[@pt] <ten or>[@en] vinte[@pt] <journals[*]>[@en] . [+ epepe]
%add: MOT (7;5.07)
%err: journals=newspapers; transference from Portuguese
Because Rafael I think he bought ten or twenty, twenty, ten or twenty journals.
F039:L31

The importance and influence of linguistic context is nicely shown in the second of these examples containing lexical errors. This time MEG is in England and talking to her father about having seen a mouse jump into a river from the riverbank. She uses the word 'banco' to refer to '(river)bank', which is incorrect as this means either 'bank' (i.e. a financial institution) or 'stool'.

(89)
*MEG: <era um banco[*] (.) que[/] que estava na[//] em frente do>[@pt] <river>[@en] <e o ratinho>[@pt] <jump(ed)>[@en][///] <err pulou na água, aí[/] aí eu vi aí uma coisa tinha pulada na água>[@pt] . [+ pepep]
%add: PAI (7;10.08)
%err: banco=margem, transference from English 'bank'
It was a bank that, that was in, in front of the river and the mouse jumped, err jumped in the water and then, then I saw then a thing had jumped into the water.
F74:L224

The Portuguese equivalent, 'margem' is either not known or not accessible at the time of speaking. The fact that she says that the 'banco' was in front of the river does imply that she feels the need to clarify what she meant by this word, thus indicating that she was not entirely happy with her lexical choice.

Again, in this third lexically-based error, it is the polysemy of a Portuguese word which results in MEG's error. Talking about a change in the time she is being picked up by a school friend the following day, after using the Portuguese 'horario', MEG uses 'moved' instead of 'changed'.

(90)
*MEG: <but the[/] the>[@en] <horário>[@pt] <has moved[*]>[@en] . [+ epe]
%add: MOT (7;11.05)
%err: moved=changed
But the, the time has moved. F078: L331

If she had continued in Portuguese, MEG would probably have used the Portuguese 'mudou' which can mean 'changed' as well as 'moved'. Her choice of 'moved' is most likely to have been reinforced by its phonetic similarity to 'mudou', thereby supressing easy access to the verb 'change', already used by MEG in other contexts on four other occasions[137]. If MEG had used the the word 'time' instead of the Portuguese 'horario' (i.e. had not code-switched here in the first place) she may have been primed to use 'changed' instead of 'moved'. However, in this instance, the school-related conversation was enough to foreground the Portuguese word for 'time', its use prompting the erroneous use of 'moved'.

The following two examples (occurring in the same dialogue) again reveal how code-switching can prime a speaker into making linguistic choices which they might not normally make when speaking monolingually. MEG is talking to her mother about her school production for which she needs a pair of black pumps. In an earlier conversation her father had suggested she should either colour (dye) some shoes she already has or wear her school shoes, so as not to buy ones just for the production. Having already talked about this school-related issue in Portuguese with her father means that MEG would clearly have to work hard to suppress any influence from Portuguese in her dialogue with her mother. In terms of actual code-switching, MEG appears to limit the insertion of Portuguese to a noun in the first

[137] A simple search for all forms of 'change' in the corpus was achieved by using the following command line: freq @ +t*MEG +s"change*" +u.

example ('festa') and a noun with an adjective ('tênis preto') in the second utterance, apparently maintaining English as the Matrix Language.

(91)
*MEG: <Daddy@p said I should paint it[*] black for the>[@en] <festa>[@pt] . [+ ep]
%add: MOT (8;1.19)
Daddy said I should paint it black for the party. F090: L41

(92)
*MEG: +" <or[*] I use[*] the>[@en] <tênis preto>[@pt] <that I use[*] for school>[@en] +...
[+ epe]
%add: MOT (8;1.19)
%err: or should be either
"Or I use the black trainers that I use for school..." F090: L48

Although the surface-level realisation of these two utterances appears accurate enough, there is clear evidence that the role of Portuguese goes beyond that of the Embedded Language. Firstly, MEG uses the singular 'it' instead of 'them' to refer to the shoes, clear influence from Portuguese where it is common to refer to (pairs of) shoes in the singular form. Support for this interpretation is found in MEG's use of 'tênis preto' (example (92)): although the Portuguese word for trainers is the same for both the singular and plural forms, the fact that she does not pluralise the adjective indicates that she is treating 'trainers' as singular. In the second example we also find two instances of 'use' marked as errors. Although one could possibly argue that 'use' is acceptable, the desired meaning here is that of 'wear', which, in Portuguese is most often expressed by the polysemic verb 'usar'. A quick KWAL analysis[138] showed that MEG was familar with the English 'wear', having used it (and its derivations) productively on 9 occasions. It is evident that MEG's choice of 'use' instead of 'wear' is strongly influenced by the orginal language of the subject matter and, in addition, by the fact that she is translating a direct quote from the father (note the quote marks coding at the beginning of her utterance) which would orginally have been in Portuguese.

The final error to be considered in example 92 is MEG's use of 'or'. On examination of this utterance in its linguistic context, we see that MEG continues her father's quote in the following utterance with +" <or they>[@en] <arranja um par de sapatilhas>[@pt]. [+ ep]. ('or they get a pair of pumps'). Whereas in English one would use 'either..or' to express alternatives, in Portuguese, this is achieved by just

[138] kwal @ +t*MEG +s"wear*" +u

repeating 'or', as in 'ou..ou'. It is likely that MEG's repetition of 'or' is a direct result of her translation of her father's original utterance. One could argue that it may also be that she is simply unfamiliar with the English construction and a search for the word 'either' in MEG's utterances[139] did indeed return zero occurences. Although this might be seen as lending support to the latter supposition, when the same search was performed on MOT's utterances[140] the output still only returned two examples and neither of these were used with 'or'. Despite there being no attested examples in the corpus we would still assume that an adult native speaker of English would be familar with the 'either...or' construction. Although the same could not be assumed for an eight-year-old child, this simple analysis illustrates how cautiously zero output must be interpreted.

Although the explanation for MEG's use of 'or' is inconclusive, there is little doubt about the influence of Portuguese in the surface realisation of the two CS utterances shown above. It is probable that transference from Portuguese into English, and vice versa, would also be found in the siblings', and indeed the adult speakers', monolingual speech. However, it is plausible to suggest, given the evidence, that by actively code-switching (i.e activating both codes), MEG is making herself more susceptible to the influence of the competing language and that this in turn means errors are more likely to occur in CS utterances. Although a thorough study of all of MEG's errors (whether in bilingual or monolingual utterances) would shed more light on the issues discussed above, such investigation would not be feasible here, as illustrated by the discussion generated by only 7 utterances!

6.3.2 Errors in JAM's code-switched utterances

The problem of having to limit exemplification to a few illustrative examples would appear to be exacerbated in JAM's case: of his 956 error codes, 180 occur in CS utterances, three times as many as MEG's. However, whereas MEG's errors proved to be mostly unique in nature (i.e the same error rarely occurring more than once), the frequency word list for JAM revealed that several of his errors were reoccurring, 'which' topping the list with 19 occurrences! In fact, the top ten words in his frequency list account for 77 (40%) of the 193 tokens coded as errors in his CS utterances. Such quantitative data is clearly very useful in guiding a subsequent qualitative

[139] kwal @ +t*MEG +s"either" +u
[140] kwal @ +t*MOT +s"either" +u

analysis of the errors and ensures a more representative selection of example utterances. Despite prioritising JAM's *reoccurring* errors in CS utterances, attention will also be given to isolated errors if they are felt to illustrate the relationship between code-switching and the production of errors.

To begin with I examined all 19 CS utterances where 'which' had been coded as an error. At times it was necessary to return to the original dialogue in the relevant file to determine why JAM's use of this word was considered to be incorrect. However, with the file name and line number available in the output, this was carried out quickly within the CLAN window itself. Such examination revealed that JAM was using 'which' as a relative pronoun to mean 'that' (13 cases), 'who' (3 cases), 'where' (2 cases) and even 'when' (1 case). The following four examples are illustrative of this generic use of 'which', the first utterance showing that JAM uses 'which' instead of 'who' to refer to a classmate who was dressed in a spiderman's outfit.

(93)
*JAM: <I saw my>[@en] <coleginha Vitor@pn>[@pt] <which[*] was>[@en] <roupa de homen+aranha>[@pt] <and then I[/] <I doesn't>[//][*] I see his eyes and then he doesn't put the>[@en] máscara[@pt] . [+ epepep]
%add: MOT (4;3.21)
I saw my classmate Vitor which was spiderman's clothes and then I doesn't, I see his eyes and then he doesn't put the mask. F010: L135

In the second example below (94), 'where' would be the appropriate relative pronoun although 'which' would be considered correct with the addition of the preposition 'in' (either before the pronoun or after 'sleep').

(94)
*JAM: <bedroom[//] the bedroom which[*] we>[@en] dorme[@pt] . [+ ep]
%add: MOT (4;4.25)
Bedroom, the bedroom which we sleep. F016: L159

In the following example 'when' would have been the best choice although 'that' would also have made grammatical sense.

(95)
*JAM: <was[*] Saturday which[*] we[//] Vincent@pn, he>[@en] ligou[@pt] <to[*] here>[@en] ? [+ epe]
%add: MOT (5;5.07)
%err: was=it was; to = influence from Portugese 'para'

Was Saturday which we, Vincent, he called to here? F071: L361

The last example below illustrates JAM's tendancy to use 'which' when either 'that' or no relative pronoun would be required in English.

(96)
*JAM: <ontem de noite eu sonhei>[@pt] <which[*]>[@en] <eu tirei>[@pt] <two[/] two teeth>[@en] . [+ pepe]
%add: MOT (5;8.00)
Last night I dreamed which I took out two, two teeth. F086: L336

The question arising here is whether JAM only uses 'which' in this way when he is code-switching or whether he also does this when speaking monolingually (in English). In order to investigate this question I used KWAL to provide me with any of JAM's monolingual utterances where he uses 'which' erroneously (i.e those coded as errors)[141]. A read through of these utterances revealed that JAM did indeed use 'which' erroneously in the same way as shown in the CS examples above. In fact out of 114 occurrences of 'which' found in his English utterances (which also include uses such as 'which one'), 43 were coded as errors. This is in contrast to MEG who only used 'which' erroneously 3 times out of a total of only 38 occurrences[142]. The difference between the siblings in overall frequency of 'which' in the corpus (151 for JAM and 45 for MEG[143]) highlights the fact that JAM is employing it as an all encompassing relative pronoun, with this overuse inevitably resulting in errors.

It is very likely that JAM's excessive use of 'which' is the result of Portuguese influence: the Portuguese relative pronoun 'que' would be perfectly acceptable in all of the four examples shown above as it can carry the meaning of 'which', 'that', 'who' and 'where'. The fact that such influence is rarely seen in MEG's utterances could mean one of two things: that JAM's usage is idiosyncratic or that this type of error is related to linguistic development and might eventually disappear as he matures in both languages. Looking at JAM's output data for 'which' from a longitudinal perspective, what we discover is that right up until the pen-ultimate recording JAM is still making the same error. This means that over the time span of the data in the corpus, which equates to JAM's age between 3;5.18 and 6;9.25, there is no evidence to suggest that he is 'growing out of' this error. Looking at MEG's data from the same

[141] Kwal @ t*JAM +s"which" +u +d -s"[+ *]" | kwal +s"[*]"
[142] Kwal @ +t*MEG +s"which" +u +d -s"[+ *]" | kwal +s"[*]"
[143] Achieved by the commands freq @ +t*JAM +s"which" +u and freq @ +t*MEG +s"which" +u

perspective what we see is that her last 'which' coded as an error was recorded when she is 6;10.00, meaning that from then on (until 9;2.19) there were no further cases of 'which' recorded as errors. It might very well be that earlier data on MEG would have shown the same frequency of errors with 'which' as shown for JAM. However, this is necessarily pure speculation as the data is not available. Equally it is not possible to state that JAM would naturally follow MEG's developmental trajectory and shortly stop producing such errors. Indeed one cannot even rule out the possibility that his use of 'which' may be slightly idiosyncratic and might not be found to such an extent in other bilingual Brazilian/English children. Evidently this would necessiate further comparative research.

Although such discussion may appear to be detracting from the original question of the relationship between code-switching and the production of errors, it has been important as a way of establishing that at times it is crucial to look beyond the data under investigation (CS utterances) to arrive at more reliable interpretations. By doing this with 'which' it has been possible to provide ample evidence to support the idea that JAM's errors when using this relative pronoun are a direct result of the underlying influence, and dominance, of Portuguese in the realisation of English surface-level grammatical morphemes. As this influence manifests itself in both monolingual English utterances and CS utterances we cannot therefore say that JAM's errors with 'which' are caused solely by his surface-level switching from one language to another.

When we examine JAM's other frequently occurring errors we find further evidence to support the idea that many of the errors in his CS utterances (and in his monolingual utterances) can be attributed to the underlying influence of Portuguese. Looking at two examples of his errors involving 'to' (15 in total), we see how the flexibility of a Portuguese preposition ('para') can have an effect on the surface-level realisation of two slightly different English structures. In the first example, talking about his Grandma changing planes on her flight back to England, JAM says 'time to she [...]' instead of 'time for her to [...]'.

(97)
*JAM: <time to[*] she[*]>[@en] <trocar de >[@pt] ? [+ ep]
%add: MOT (4;10.00)
Time to she change? F032: L1118

Although rather clumsy in English, the Portuguese translation 'hora para ela' would be felictious, indicating that Portuguese might be affecting how JAM realizes the English structure. Of course if JAM had not used a pronoun ('she' instead of the correct 'her' in this case), his use of 'to', as in 'time to change', would then have been correct: it is the insertion of an object pronoun which necessitates the change from 'to' to 'for' in English. In Portuguese there is no need for a change, 'para' being used in both structures.

The second example reinforces the idea that JAM is being influenced by this more flexible use of 'para' when he says '[...] room to we sleep [...]' instead of 'room for us to sleep':

(98)
*JAM: <loads of room to[*] we[*] sleep and the>[@en] <piscina funda>[@pt] <and a tiny one>[@en] . [+ epe]
%add: MOT (5;6.19)
Loads of room to we sleep and the deep swimming pool and a tiny one.
F079: L236

In both cases, the fact that JAM uses inappropriate pronouns in English, 'she' instead of 'her' and 'we' instead of 'us', lends support to this interpretation of underlying Portuguese influence: unlike English, most Portuguese subject and object pronouns have the same form. Interestingly, a quick search of JAM's use of 'for' in the corpus[144] revealed that he did use it (39 occurrences) but mostly with (correct) object pronouns (e.g.'for me')(21 occurrences) or in the question type 'What is xxx for?" (7 occurrences). There were no occurrences of 'for' followed by an object pronoun *and* a verb.

Although it is not feasible to present JAM's remaining 13 errors with 'to', the two examples above illustrate well how the underlying cause for such errors is related to influence from Portuguese rather than being triggered by the action of code-switching itself.

With regards to the third and fourth most frequently occurring errors in JAM's word list, 'is' (9 occurrences) and 'it' (7 occurrences), what we find is that on examination of the utterances themselves these particular frequencies are rather misleading. Looking at the first example below it is clear that JAM's error involves missing out the pronoun 'it', likely influence from Portuguese where no subject

[144] combo @ +t*JAM +s"for"

pronoun is needed. In the absence of the pronoun the error code was placed next to 'is' in order to flag up the error. However, this method then means that a frequency analysis will compute 'is' as the error and not the absence of 'it'.

(99)
*JAM: <look, is[*] a>[@en] mosquinha[@pt] . [+ ep]
%add: MOT (4;5.00)
%err: no "it"
Look, is a little fly. F017: L129

The correct form of coding, as suggested in CHAT, would be to insert the missing word and precede it by a zero to indicate that it was missing (0it[*]). The CS utterances in the output for JAM showed that I had not been consistent in following this suggestion: out of 14 errors which involved JAM missing out a word, only two had been coded correctly. The 12 missing words were subsequently seen to be as follows: 'it' (4 occurrences), 'am' (2 occurrences), 'you' (2 occurrences) and 'I', 'to', 'is', 'are' (each with one occurrence). With regards to the original frequency for 'is' and 'it' (9 and 7 occurrences respectively), the correct coding would then result in 7 errors involving 'is' and 11 involving 'it'. If we consider that only 12 out of a total of 180 error codes were coded erroneously and that the resulting differences in frequency were not greatly affected, this methodological issue is not a cause for alarm for this particular study (although corrections to the coding will be made). However, for spoken data where the need for accurate quantitative analysis of errors is important (such as in studies of younger children, those with speech impairments or second language learners) it is evident that an effective and consistent coding system is paramount.

Returning to JAM's errors with 'is' and 'it', while example (67) above illustrates his missing out of the pronoun, the example below (with the original error coding) shows him missing out the auxilary 'is' and 'are':

(100)
*JAM: o(lha)[@pt] <(other)wise she[*] going to>[@en] pensar[@pt] <which[*] we[*] at home, yeah[@tq]>[@en] ? [+ pepe]
%add: MOT (5;3.10)
%err: misses out "is"; uses "which", transference from Portuguese; misses out "are"
Look, otherwise she going to think which we at home yeah? F052: L32

It is unlikely that JAM's use of 'olha' and 'pensar' in this CS utterance has caused JAM to miss out the two forms of 'be'. Indeed an analysis of all of the 156 cases where he

uses 'going' or 'gonna'[145] (mostly in monolingual English utterances) reveals that in the data up to file 092 (when JAM is aged 5;9.06) he is seen to miss out the auxilary 'am' (12 times), 'are' (11 times) and 'is' (4 times). The fact that 22 of these 27 errors occur in the first half of the data (up to file 058) indicate that this error is related to his linguistic development, there being proportionately more correct auxilary insertion as his language matures. This interpretation is supported by the same anaysis of MEG's data[146]: of the 181 times she uses 'going' or 'gonna' there are no missing auxilaries. If one accepts that correct auxilary usage is age-related one should expect JAM to soon outgrow this type of error, thereby following the linguistic trajectory of his sister.

These additional analyses have been important in showing that while some of JAM's errors can be attributed to the underlying influence of Portuguese, in the case of the missing auxiliaries (with 'going') discussed above, it seems more likely that they are simply developmental errors unrelated to the use of another language. This ultimately means that the act of code-switching is not responsible for their occurrence.

Further analyses of JAM's errors involving 'it' in CS utterances reveal that apart from missing the pronoun out (due to Portuguese influence), on several occasions (9 in total) he uses it erroneously as part of a tag question. For example, in (101) JAM uses 'isn't it' instead of 'isn't there':

(101)
*JAM: <depois de amanhã>[@pt] <there is>[@en] <escola>[@pt], <<isn't it>[@tq] [*]>[@en] ? [+ pepe]
%add: MOT (5;7.24)
After tomorrow there is school, isn't it? F085: L318

Although in this case it is only the 'it' which is at fault, in the majority of cases both parts of the tag question are erroneous, as shown in this second example:

(102)
*JAM: <mas tu[/] tu (es)tá um>[@pt] <invisible one <isn't it[*]>[@tq]>[@en] ? [+ pe]
%add: MEG (5;9.06)
But you, you are an invisible one, isn't it? F092: L370

[145] The command combo @ +t*JAM +s"going" +u automatically captures cases of 'gonna' as these were transcribed as 'gonna[: going to]', thereby allowing COMBO to include them in the search.
[146] combo @ +t*MEG +s"going" +u

Correct question tag usage here would have resulted in 'aren't you'. In total, of the nine incorrect uses of 'it' in tag questions, seven are combined with 'isn't', one with 'is' and one with 'was'. This high frequency of occurrence of 'isn't it' is not exclusive to JAM's *CS* utterances, as revealed by a further analysis which also looked at his monolingual English utterances[147]. On 15 occasions JAM used 'isn't it' (11 times) and 'is it' (4 times) incorrectly, i.e when different combinations of auxiliary verbs and pronouns were necessary. If we add together the incorrect usage found in CS utterances and monolingual English utterances we arrive at 24 errors involving 'isn't it', 'is it' and 'was it'. In terms of proportions what we then find is that while 62% of these errors occur in his monolingual utterances, 38% are found in JAM's CS utterances. Considering that JAM's overall number of CS tokens represents only 17% of his total word count, the percentage of tag question errors involving 'it' in CS utterances (38%) seems to be disproportionately high. This is an indication that JAM's CS utterances appear to be more prone to this type of error than his only English utterances. Although it would be fitting to now suggest reasons as to why the latter occurs, this discussion will be taken up in the next section which focusses solely on the siblings' use of tag questions in their CS utterances.

Having already examined 10 examples of JAM's most frequently occurring errors in CS utterances I will now briefly look at two more errors which, although more infrequent, prove to be quite telling about how JAM attempts to combine his two languages in a single utterance.

In this penultimate example JAM is at the airport seeing his Grandma off and is asking why the wind (air) that comes out of the plane is so hot. Although the word 'the' has been coded as an error (twice) the problem here lies in the apparent absence of the word 'is' after 'why'.

(103)
*JAM: <Mummy@m, <why the>[/] why the[*]>[@en] vento[@pt][//] <why the[*]>[@en]
<vento é[/], é muito quente>[@pt] ? [+ epep]
%add: MOT (4;10.0)
Mummy, why the, why the wind, why the wind is, is very hot? F032: L778

However, when we look at the Portuguese contribution to the utterance we do in fact find the word 'is', 'é', repeated twice. JAM appears to be favouring the grammatical structure of Portuguese where the inversion of subject and verb or the insertion of an

[147] Kwal @ +t*JAM +u +s"[*] +d | kwal +s"it"

auxiliary is unnecessary after question words. The fact that he does this in two other CS utterances ('Why I can't tomar picolé?' (F034: L624) and 'Why I am pelado' (F077: L294) indeed suggests that Portuguese is having an influence on the surface-level realisation of his English morphemes. Even in his monolingual English utterances[148] JAM rarely places the auxiliary immediately after 'why'. However, as we investigate further and take a look at the data for 'why' from a longitudinal perspective this interpretation (that JAM's lack of subject auxiliary inversion after 'why' is due to the influence of Portuguese) is called into question. After the age of 5;7.24 (File 085) there are no more errors recorded: all 8 occurrences from this age on show correct inversion. It is difficult to say whether this correct usage is due to JAM's knowledge of English structure maturing or to a diminishing influence of Portuguese. The fact that no such errors with 'why' can be found in any of MEG's utterances[149] supports the first supposition, but only if we consider the siblings to be following the same developmental pattern in English.

Again, this discussion has shown the importance of being able to look beyond the particular data under analysis (the errors occurring in the siblings' CS utterances) in order to avoid erroneous interpretations. Such subsequent analyses (involving searches for the same errors in monolingual utterances) are greatly facilitated by the methodology of this study where the search and retrival of data is instantaneous. It is also relevant to highlight here how the longitudinal nature of the LOBILL Corpus itself means that these analyses are particularly enriching: we are able to track the siblings' errors over time and take into account developmental aspects, something which would not be possible in a synchronic corpus.

By looking at the overall KWAL output of JAM's errors in CS utterances from a longitudinal perspective it is possible to see that there is a point when he appears to cease producing errors in his CS utterances. In Files 099 to 119, which correspond to JAM's age between 6;3.08 and 6;9.25, there is in fact only one recorded error, in File 116 (see below). It is likely that this drastic reduction of errors is due to his linguistic development, especially in English, with his increasing proficiency clearly aided by the family's move to England (all the recordings from File 100 onwards took place in England). It is this increasing dominance of English in JAM's linguistic environment that explains his last attested CS error, discussed below.

[148] Kwal @ +t*JAM +s"why" +u +d
[149] Kwal @ +t*MEG +s"why" +u +d

Five months are having arrived in England JAM produces the following error in a CS utterance addressed to his Brazilian grandfather over the phone. He is trying to explain that although he is still going to be seven, he is already in the class where children are seven:

(104)
*JAM: <eu sei mas eu (e)stou[//] eu[/] eu[/] eu vou ser[*] sete>[@pt] <n(ext)[@en]>[?][//] <estou sete mas, vai[/] vai ser um[//] (por)que eu estou na classe de fazer sete>[@pt] . [+ pep]
%add: AVO (6;8.05)
I know but I am, I, I, I am going to be seven next, I am seven but it's going, going to be one, because I am in the class of becoming seven. F116: L702

In terms of code-switching, JAM inserts the word 'next' (or rather begins to say it before retracing) into an otherwise monolingual Portuguese utterance. His various retracings and repetitions (as evidenced by the symbols [//] and [/]) show that he is clearly struggling to express what he wants to say in Portuguese. This is understandable given that apart from his immersion in an English context, he is talking about a (British) school-related issue. Although JAM realises that he cannot revert to English due to his addressee's monolingualism, he struggles to suppress its underlying influence: instead of saying 'vou *fazer* sete anos' ('I'm going to *do* seven years old') or possibly 'vou *ter* sete anos' (I'm going to *have* seven years old'), he translates from the English and uses the verb 'to be'. As with other romance languages such as Spanish and French, in Portuguese, age can only be expressed with the verb 'have' (or 'fazer' ('do') in the case of Portuguese when there is future reference) and the use of the equivalent of 'years old' is almost obligatory. Here, JAM's Portuguese is clearly being influenced by his English[150] - this is in stark contrast to what we have seen up to now in this section. Indeed, over a year previous to this CS utterance, the influence was clearly the other way round, causing JAM to produce errors such as '<Jake@pn <he's got>[*] five years old>[@en]' (F055: L132)[151]. On one occasion, perhaps to avoid such an error, JAM is even seen to switch to Portuguese: '[...] when[/] when I>[@en] <tinha quatro anos>[@pt] ? [+ ep]'(F078: L188)[152], using the past form of 'ter' ('have'). Occurring in a conversation with his bilingual mother, JAM is able to make use of the code-switch to facilitate expression and avoid

[150] Although JAM also uses 'estou' (I am) three times in this utterance, his last use 'estou na classe de fazer sete' is correct and implies that this might have been his intention with the other two uses of 'estou'. For this reason they are not coded as errors.
[151] Obtained by the following search: kwal @ +t*JAM +s"got" +u +d
[152] Obtained by the following search: kwal @ +t*JAM +s"anos" +u +d

a potential error. As can be seen in example (104) presented above, when addressing a monolingual speaker, JAM cannot make use of such a code-switching strategy and this ultimately results in the error.

From the discussion of this last example (and related corpus searches) an important point has now emerged regarding the relationship of code-switching and the production of errors. Although the frequency analyses have shown that relatively more errors tend to occur when the siblings are engaged in code-switching (rather than speaking monolingually), we must also consider the fact that JAM and MEG may actually be avoiding further potential errors by code-switching. By inserting a particular word or expression from the other language, potential grammatical or lexical conflicts arising from the use of non translation equivalents is avoided. Presently it is not possible to investigate this matter further to try and determine how often JAM and MEG use code-switching as a strategy to avoid errors. All that can be said is that if a total of 180 errors have been noted in JAM's 656 CS utterances[153] , this means that at least 476 of these utterances are well formed. It is likely that this number is actually greater if one considers that some utterances could contain more than one error (increasing therefore the number of correct utterances). And with regards to MEG's CS utterances, with only 48 errors in total, at least 334 of her 382 CS utterances[154] are well formed.

Of course, one must bear in mind that these figures are dependent on the accuracy and consistency of error coding in the corpus, which in turn implies that the coder has an understanding (based on a theory) of what constitutes an error, whether in English or Portuguese. Although such issues with reliability can be addressed by using more than one coder, the limitations of this particular study meant that this was not viable. Despite this caveat, however, the discussion of the siblings' errors in CS utterances in this section has been extremely fruitful in shedding light on how complex the interplay of English and Portuguese can be in bilingual, and even monolingual, speech. The analyses have shown that, particularly where JAM is concerned, the influence of Portuguese can be held accountable for many of the errors found in the CS utterances. However, as was seen through the longitudinal analyses of the data, both developmental and contextual aspects clearly have important parts to play in explaining the production, and decrease in production,

[153] This total was obtained by the following search: freq @ +t*JAM +s"<+ *>" +u
[154] Obtained by the search freq @ +t*MEG +s"<+ *>" +u

of the siblings's errors. This is especially true for JAM, whose increasing competence in English, accelerated by his move to England, appears to help re-address the apparent underlying dominance of his Portuguese.

Although comparisons between JAM and MEG have been made in this discussion, such as when investigating the developmental trajectory of certain errors (the use of 'which', the missing out of the auxilary before 'going' and the lack of inversion after 'why'), only a more indepth comparative study of the siblings' errors (in both bilingual and monolingual utterances) would provide more adequate data from which conclusions could be drawn regarding the relative influence of the factors mentioned above. Returning to the initial question of the relationship between the production of errors and code-switching, it has become evident from the examination of the few examples in this section that this is a complex issue which cannot be fully resolved here. However, there was one particular type of error of JAM's, involving the use of 'it' in tag questions, which was noted as being particularly frequent in his CS utterances (when compared to his monolingual utterances) and indeed, as will be seen, appears to be caused by the act of code-switching itself. It is to this discussion on tag questions which I will now turn.

6.4 An utterance-level analysis of tag questions in code-switched speech

In section 5.2.4, quantitative analyses of the tag question code ([@tq]) revealed that JAM used tag questions much more frequently than MEG and that a high proportion of these appeared in his CS utterances: 32 out of JAM's 119 tag questions occurred in CS utterances while out of MEG's total of 48 only 2 tag questions occurred in her CS utterances. Having used KWAL to locate these utterances (along with the addressees) for both JAM[155] and MEG[156], the aim of this section is to examine these tag questions in their linguistic context to determine how, and why, they are used in bilingual utterances. As the data will be examined from a longitudinal perspective, the ages of the siblings have been included (in brackets) in the examples discussed.

6.4.1 Tag questions in MEG's code-switched utterances

Beginning with the two occurrences in the data for MEG, what we see in both cases is the generic 'yeah' being attached to an otherwise Portuguese utterance:

[155] kwal @ +t*JAM +s"[@tq]" +u +s"[+ *]" +t%add +fJAMtq
[156] kwal @ +t*MEG +s"[@tq]" +u +s"[+ *]" +t%add +fMEGtq

(105)
*MEG: <a vez da Mamãe@m>[@pt] <yeah[@tq]>[@en] ? [+ pe]
%add: JAM MOT (7;9.10)
Mum's turn, yeah? F063: L1158

(106)
*MEG: <é cara de pau>[@pt] <yeah[@tq]>[@en] ? [+ pe]
%add: MOT (8;6.29)
That's cheeky, yeah? F097: L709

In the first example, the siblings and their mother are playing a game of mini-snooker and MEG is simply checking whose turn it is. The use of 'yeah?' instead of the Portuguese tag question 'é?' ('is it?') appears to be a cursory nod to the mother seeing that the utterance itself seems to be more directed at her brother. By using this generic tag question, MEG avoids any potential mismatch between the Portuguese part of the utterance and an English canonical tag question. This also appears to be true of the second example, where MEG is accusing her mother of being cheeky when the latter reminds her daughter that Mother's Day is coming up and she expects a treat. A quick frequency analysis of this particular meal time conversation[157] showed that 80% of the tokens MEG addresses to her mother are actually in Portuguese. It may be that MEG's use of 'yeah' here is a token gesture to compensate for the use of so much Portuguese with her mother in this conversation. What is certain is that she does not switch to English because she is unable to recall the Portuguese tag question 'é' or 'né': of the 46 tag questions she uses in monolingual (Portuguese or English) utterances[158], 'né' accounts for 21 occurrences and there are two cases of 'é'.

Although in the two CS utterances above MEG appears to avoid using an English canonical tag question, an analysis of the 46 non-CS utterances shows that she does actually use them, correctly, in monolingual English utterances: amongst the 18 English tag questions she uses we find the following combinations: 'isn't it' (6 occurrences), 'isn't he' (2), 'didn't it' (2), 'didn't you' (1), 'wasn't it' (1), 'don't you' (1), 'can't we' (1) and 'is it' (1). However, it is evident that instead of trying to add an English

[157] This was achieved by first selecting file 097 and then using the commands kwal @ +t*MEG +t%add +s"MOT" +d | freq +s"<@pt>" and kwal @ +t*MEG +t%add +s"MOT" +d | freq +s"<@en>" in order to compare the contribution of both languages in terms of overall tokens.
[158] Kwal @ +t*MEG +u +s"[@tq]" -s"[+ *]" +t%add +fMEGtq

canonical tag question to a Portuguese utterance, MEG's strategy has been to go generic (with 'yeah'), thus avoiding potential errors.

From the evidence already presented in the last section on JAM's errors involving 'it', it appears that such an avoidance strategy is not taken advantage of by JAM when using tag questions in CS utterances. A complete analysis of his tag question usage will shed more light on this apparent difference between the siblings and reveal more about how this younger sibling makes use of tag questions in both bilingual and monolingual utterances.

6.4.2 Tag questions in JAM's code-switched utterances

With 32 instances of tag question usage in CS utterances (and 87 in monolingual utterances) it seemed viable to examine JAM's utterances from a longitudinal perspective. By doing this I would be able to track particular tag questions, such as 'isn't it', which JAM frequently appears to use incorrectly. As comparison with such usage in monolingual utterances would benefit this longitudinal analysis, I will be making comparisons throughout the discussion rather than first presenting just the CS data. However, for reasons of space it will not be possible to present examples from the non-CS data.

In the first 30 files of the corpus there is only one instance of JAM using a tag question in a CS utterance, as shown below when he adds a 'yes' to the utterance after a switch to Portuguese:

(107)
*JAM: <he's going home to eat his>[@en] papapa[@pt], <yes[@tq]>[@en] ? [+ epe]
%add: MOT (4;5.24)
He's going home to eat his food, yes? F023: L39

During the same period, in JAM's non-CS utterances (monolingual utterances), there are 7 tag questions in total: 2 in Portuguese (both 'né') and 5 in English (2 'yes', 2 'is it' and 1 'no'). Significantly, one occurrence of 'is it' (F022: L513) is incorrect, revealing that JAM has made an error with a canonical tag question when speaking monolingual English.

If we now look at Files 031 to 052 which cover the time period between 31st December 2002 and 11th June 2003, which was just before the siblings' trip to England, we see the frequency of JAM's tag questions increase: to 18 occurrences in

CS utterances and 16 in non-CS utterances. Considering that only 17% of JAM's total word count is found in CS utterances, the frequency of his tag questions in CS utterances is much higher, proportionately, than in his non-CS utterances. Despite this difference in comparative frequency, the types of tag questions are remarkably similar: in his CS utterances JAM uses 'yeah' 12 times, 'isn't it' 5 times and only one Portuguese tag question 'tem' ('has'); in his non-CS utterances he uses 'yeah/yes' 12 times, 'isn't it' 3 times and the Portuguese tag question 'vai' ('going') once. What one then notices (due to the error coding) is that every single 'isn't it' (8 in total) constitutes an error! These errors result from the blanket use of 'isn't it' where different combinations of auxiliaries and pronouns would normally be used, as can be seen in the example below:

(108)
*JAM: <(be)cause she[/] Meggie@pn>[@en] <todo dia>[@pt] <she want[*] to be the first, <isn't it>[@tq][*]>[@en] ? [+ epe]
%add: MOT (5;0.03)
Because she, Meggie, every day she want to be the first, isn't it? F038: L:1510

A simple explanation for this all-encompassing use of 'isn't it' is transference from the generic Portuguese tag question 'né', an abbreviation of 'não é', which translates as 'isn't it'. An alternative explanation might be that JAM has still to develop the more complex English canonical system of tag questions. As we continue to examine his utterances diachronically it is hoped that the evidence will provide further support for either, or both, explanations.

The next two months of recordings (Files 053 to 075) took place in England where the siblings were on holiday with their mother, visiting English relatives. During this period there is only one occurrence of a tag question in JAM's CS utterances, the Portuguese tag 'não é'. This occurs in a telephone call to his father in Brazil. However, in his monolingual utterances there are 14 occurrences, of which 9 are English tag questions ('yeah' and 'isn't it') and 5 are Portuguese ('é' and 'né'). It is interesting to note that in contrast to the previous period, the generic tag 'yeah' now only occurs twice and 'isn't it' is used 7 times. In terms of Portuguese tags we also see more use of 'né', occurring 4 times, and 'é', occurring just once. Added together, the equivalents 'isn't it' and 'né' account for 11 of all the 14 tag question occurrences in monolingual utterances. It is likely that JAM's increase in the use of 'isn't it' is related to his

increase in the use of the Portuguese equivalent 'né'. This interpretation is given more support by the fact that 6 of his uses of 'isn't it' are marked as errors, JAM appearing to simply map the grammar of the Portuguese generic tag question onto his English utterances.

It is only in the fouth period under study that we find occurrences of English tag questions not previously used by JAM. Covering the 10 months following his return to Brazil from England (Files 076 to 099), Jam is recorded using 'was it', and 'is he' in two CS utterances and 'was it', wasn't it, 'did it' and 'would it' in four of his monolingual English utterances. It is possible that two months of intensive English input has had something to do with his increase in variety of tag question usage. However, as we can see from the two CS examples, shown below, JAM's canonical system still needs to develop somewhat. In this first example JAM is now able to match the auxiliary of the tag question with the verb of the main sentence despite a single word code-switch to Portuguese, which could have triggered a Portuguese tag.

(109)
*JAM: <I was the size of a>[@en] <formiga>[@pt], <<was it[*]>[@tq]>[@en] ? [+ epe]
%add: MOT (5;6.01)
I was the size of an ant was it? F077: L271

However, instead of also matching the pronoun 'I', JAM uses 'it'. This appears to be the opposite of what occurs in the second example below. Here we see the correct use of the pronoun 'he' but the use of 'is' instead of a more appropriate 'does' ('or doesn't').

(110)
*JAM: <<but they[/] they>[//] but he>[@en] <fala em português>[@pt] <as well, <is[*] he>[@tq]>[@en] ? [+ epe]
%add: MOT (5;8.00)
But they, they, but he speaks in Portuguese as well, is he? F086: L249

It is perhaps not surprising that JAM makes an error here given that the original verb is in Portuguese. However, even if he had not code-switched to Portuguese it is unlikely that JAM would have produced 'does' as there has been no evidence of such an occurrence in the data thus far examined. Of the four cases of novel tag questions in his English monolingual utterances, while 'was it' was used erroneously, the other three ('wasn't it', 'did it' and 'would it') were used correctly.

Looking now at JAM's use of the ubiquitous 'isn't it' during this period, there are 7 instances in the CS data and 7 in the non-CS data. Compared to the previous periods, JAM now appears to be relatively more successful in his usage of 'isn't it': 3 of these tags are used correctly in CS utterances while 4 are used correctly in non-CS utterances. However, it is likely that his correct usage is coincidental rather than proposital, as the example below indicates:

(111)
*JAM: <não esse aí é o ultimo>[@pt] <<isn't it>[@tq]>[@en] ? [+ pe]
%add: MEG MOT (5;8.25)
No, that one there is the last, isn't it? F091: L138

With the main part of the utterance in Portuguese, JAM would first need to be able to recognise the verb and subject, translate them into English and then work out what tag question he should use. Rather than having gone through this complex process it is much more likely that JAM automatically uses his standard English tag question which, in this case, turns out to be harmonious with the main utterance. This interpretation is given further support when we examine the following utterance which JAM addressed to his mother 10 days before moving to England in June 2004 when he was 6;3.07:

(112)
*JAM: <(be)cause I[/] I think which[*] Jake@pn and Max@pn got that beyblade>[@en]
<<né>[@tq]>[@pt][//] <<isn't it[*]>[@tq]>[@en] ? [+ epe]
%add: MOT (6;3.07)
Because I, I think which Jake and Max got that beyblade isn't it, isn't it?
F098: L59

Chatting about his English cousins and their beyblade, JAM uses the Portuguese tag question 'né' before reformulating and providing the equivalent English tag 'isn't it'. Although the correct tag would have been 'didn't they' there appears to be no realisation on JAM's part that anything other than 'isn't it' would be necessary. Of course, having first used the Portuguese 'né' here, JAM is further primed to simply insert the English equivalent. This is a particularly revealing example because it provides clear evidence that JAM must indeed be simply transfering from the Portuguese tag whenever he uses 'isn't it'.

Before moving on to the last period which covers the six months following the siblings' arrival in England, it is important to highlight that during this fourth period the frequency with which JAM uses tag questions in monolingual utterances surpasses that of his CS utterances: whereas in his CS utterances there are 12 occurrences of tag questions (10 English and 2 Portuguese), a total of 29 tags can be found in his non-CS data (22 English and 7 Portuguese). Of these 29 tags, 12 involve the use of 'isn't it' (7 occurrences) and 'né' (5 occurences), further evidence of JAM's ubiquitous use of these two translation equivalents in both CS and monolingual utterances.

When examining the results for the last period (Files 100-119), we do not find a single occurrence of a tag question in JAM's CS utterances. We do find, however, 22 tag questions in his monolingual utterances, 7 English ones and 15 Portuguese ones. Although 'yes/yeah' account for 3 of the occurrences, there are two occurrences of novel English tag questions, 'have we' and 'don't you', both used correctly: this indicates that his system appears to be maturing. It is also significant that his two uses of 'isn't it' are correctly employed. It is very likely that JAM's immersion in an English-speaking environment has resulted in his English tag question system becoming more varied and more accurate with an increasing realisation that 'isn't it' cannot be used as a generic tag (as it can in Portuguese). With regards to JAM's 15 Portuguese tag questions we also find slightly more variety, the following five occurring (in decreasing order of frequency): 'né/não é' (7), 'é' (3), 'está' (2), 'não tem' (2) and 'não' (1).

In terms of the overall frequency of tag questions in monolingual utterances for this period, it is not surprising that there are more Portuguese question tags (15) than English tags (7): this corresponds to the increased proportion of JAM's Portuguese utterances in this period when compared to the number of his English utterances (851 as opposed to 555[159]).[160] In all of the other time periods JAM's Portuguese monolingual utterances only accounted for approximately 26% of the total number of utterances. It is in terms of *types* of Portuguese tag questions, and not tokens, where we see something interesting occurring: JAM chooses to use the tag 'não tem' twice, despite the fact that the all-encompassing 'né' could easily have been used in both

[159] The total numbers for the Portuguese and English utterances for each time period were arrived at by selecting the files for the corresponding time periods separately and repeating the following two analyses: kwal @ +t*JAM +s"[@pt]" +u +d -s"[+ *]" | freq +s"[@pt]" and kwal @ +t*JAM +s"[@en]" +u +d -s"[+ *]" | freq +s"[@en]"

[160] Rather than signifying that JAM was speaking more Portuguese, this is merely a reflection of the nature of the majority of the recordings for this period - telephone calls to Brazilian relatives.

cases. It is possible to suppose that this is evidence that JAM's increasing knowledge of the English canonical tag question system is now having an influence on his Portuguese system: his tag question matches the verb of his main sentence ('tem').

It has been important to track JAM's use of tag questions in monolingual utterances as well as in his CS utterances. From the analyses discussed above it is now possible to suggest that there is a relationship between the development of his English tag question system and the types and frequency of tag questions in bilingual utterances. The evidence appears to suggest that the more accurate and more varied his tag question usage becomes in his English monolingual utterances, the less frequently we see them being used in his CS utterances. In both types of utterances, we see the tag 'yes/yeah' becoming less frequent as 'isn't it' takes over as the most popular alternative. As time progresses (period 4) we see more accurate use of this tag and a little more variety in the use of other tags, but the latter is most noticeable in JAM's monolingual utterances (English and Portuguese) where there are 10 different types in a total of 29 occurrences as opposed to only 5 types in a total of 12 occurrences in his CS utterances. While JAM's increasing accuracy and variety in tag question usage continues to manifest itself in his monolingual utterances in period 5, JAM ceases to use tag questions at all in his CS utterances. It may be that having finally grasped how complex the English tag question system is, JAM no longer feels confident in trying to use it in bilingual utterances where you may have the added complication of matching an English auxiliary and pronoun with the corresponding Portuguese verb and pronoun. Although JAM also has the option of using the generic Portuguese tag question 'né' with any English subject and verb combination, there are no ocurrences of this usage in his CS utterances in period 5 (but note 5 occurrences of 'né' in his monolingual Portuguese utterances). The notion that JAM ceases to use tag questions in his CS utterances due to self-monitoring is more plausible if we recall that the two occurrences of the generic tag question 'yeah' in his sister's CS utterances were interpreted as being a possible strategy to avoid making a grammatical error (such as the mismatch of the subject and verb in the main clause and the tag question).

Having now examined the data for JAM longitudinally, I am also in a position to comment on the finding (from earlier frequency analyses) that the occurrence of tag questions is proportionately higher in his CS utterances (32 out of 119 occurrences) than in his monolingual utterances. In order to do this more effectively I

have summarised JAM's frequency results in the table below. As seen in the discussion above (and now shown in Table 27) JAM's use of tag questions in CS utterances is far from consistent over time (see row 4): in period 1 there is a single occurrence, in period 2 there are 18 occurrences, in period 3 there is a single occurence, in period 4 there are 12 occurrences and in the final period there are no occurrences at all.

Table 27. Summary of JAM's tag question (TQ) frequency results per time period

Period	1	2	3	4	5
Files: Age:	001-030 3;5.18-4;9.24	031-052 4;9.30-5;3.10	053-075 5;3.12-5;5.16	076-099 5;5.23-6;3.8	100-119 6;3.18-6;9.25
Location of recordings	Brazil	Brazil	England	Brazil	England
Number of TQs in CS utterances	1	18	1	12	0
Number of TQ in monolingual utterances	7	16	14	29	22
Total number of CS utterances	161	153	130	159	53

This inconsistency does not appear to correlate with his use of tag questions in monolingual utterances (row 5). In fact from period 3 and on the comparative difference in numbers of tag questions occurring in CS and in monolingual utterances increases until in period 5 where there are 22 occurrences in monolingual utterances but not a single one in JAM's CS utterances. Although location does seem to have an effect on overall frequency of occurrence of tag questions in monolingual utterances (compare periods 4 and 5, row 5), this effect appears to be much more marked in CS utterances (compare periods 2 to 5, row 4). By also including information about the overall number of CS utterances for each period (row 6), we are able to see that the inconsistency in frequency of tag questions in CS utterances cannot be fully attributed to changes in the total number of CS utterances per period. Although we might expect fewer tag questions to occur in period 5 due to the reduced number of CS utterances (53), this does not explain why we should still then see only one occurrence in period 3 when JAM produces many more CS utterances (130).

It is evident that only by examining the data from a longitudinal perspective and triangulating the results with other frequency data can we hope to explain the

original finding that JAM uses more tag questions when code-switching than when speaking monolingually. What is now clear is that the vast majority of these CS occurrences (30 out of the total of 32 tag questions) can be found in two of the periods, 2 and 4, suggesting two surges in tag question usage. The first surge appears to be related to the actual discovery of the pragmatics of tag questions, leading to experimentation and the indiscriminate use of English generic tags, clearly influenced by Portuguese. The second surge is characterized by a little more variety in tag question usage and more appropriate usage (influenced by JAM's stay in England). However, with the development of his canonical tag question system, we eventually see a drastic reduction in his tag question usage in CS utterances (period 5), most likely due to increased self-monitoring. Based on MEG's tag question frequency results it is probable that JAM would continue to avoid using tag questions in his CS utterances. However, this hypothesis cannot be tested as the data is not available.

As has been seen in other sections, there are many factors to be considered when searching for explanations for the linguistic phenomena being observed in the siblings' CS utterances. This section has particularly highlighted how important a role both linguistic development and linguistic context have to play in affecting their code-switching behaviour. Of course, it is only possible to take these factors into account due to the longitudinal nature of the LOBILL Corpus and the variety of contexts in which the data was collected. It is also worth reiterating here that such analyses and discussions would not be possible if it were not for the specific coding of the corpus.

6.5 An utterance-level analysis of metalinguistic codes in code-switched speech

In section 5.2.5 the frequency analyses of the metalinguistic code ["] revealed that for both MEG and her parents, but not for JAM, there was significantly more use of this quoting device in their code-switched utterances than in their monolingual utterances. They also revealed that while for the siblings there were roughly equal numbers of English and Portuguese tokens coded with ["] in their top 20 word lists, for MOT all but one of the tokens were Portuguese words. The aim in this section is to examine the CS utterances containing these codes in order to understand how they are being used and why such differences between the speakers exist. Again, due to the limited ouput for PAI, the analysis of his metalinguistic usage will be incorporated into section 7.4 where I examine all of his code-switched utterances.

For three of the speakers (JAM, MEG and MOT) KWAL was used to output all of the CS utterances containing ["] and these were saved as separate output files[161]. As for previous analyses, the addressee string **+t%add** was also added to the command line so that this information would automatically be included in the output. I will begin by presenting the results for JAM before discussing those pertaining to MEG and MOT.

6.5.1 Metalinguistic usage in JAM's code-switched utterances

With a total of 39 occurrences of the symbol ["] in JAM's code-switched utterances, what we find is that 15 of them are being used to mark a quote, as shown in the following three examples. In the first one JAM is telling his mother about a story he was told at school that day and code-switches to Portuguese in order to quote the original words used by the teacher:

(113)
*JAM: <then they said>[@en] <<e acordou>["]>[@pt] . [+ ep]
%add: MOT
Then they said "and he woke up". F042: L89

Example (82) shows JAM quoting an English car noise to illustrate to his father over the phone what happened when he went for a ride in a real fire engine in England. The quoting of noises accounts for 6 of all quoting occurrences.

(114)
*JAM: <e a gente andou e alguem no carro+de+bombeiro que quando estava dirigindo aí alguem>[@pt] <beep+beep["]>[@en] carro+de+bombeiro[@pt] . [+ pep]
%add: PAI
And we went along and someone in the fire engine who when he was driving then someone "beep beep" fire engine. F062: L182

The third example occurs when JAM is again talking over the telephone to his father, this time reporting what he had heard his sister say about him. Interestingly JAM chooses to remain faithful to MEG's original utterance which involved a code-switch, perhaps in order to maintain MEG's apparently marked use of 'crazy' and not the Portuguese equivalent 'doido'.

(115)

[161] For example, kwal @ +t%JAM +u +s"[+ *]" +s'["]' +t%add +fJAMmeta

*JAM: +" <<o James@pn é tão>[@pt] crazy[@en]>["] . [+ pe]
%add: PAI
"James is so crazy" F104: L176

The remainder of the ["] codes found in JAM's CS utterances (24) are all related to metalinguistic usage where JAM either wants to check the meaning of a word with his interlocutor (first example shown below) or requests assistance in translating a word (second example below):

(116)
*JAM: <tu sabe que é um>[@pt] <tram["]>[@en] ? [+ pe]
%add: PAI
Do you know what a "tram" is? F064: L49

(117)
*JAM: <how do you say>[@en] <ratinhos["]>[@pt] ? [+ ep]
%add: MOT
How do you say "mice"? F074: L172

Although these utterances show that JAM knows that the words he is referring to metalinguistically come from the 'other language', when it comes to referring to the languages themselves there are occasions when he shows uncertainty as to how to label which one is which. This is clearly shown in the example below which occurs in a conversation where JAM's mother is asking him how to say certain English words ('sand', 'beach' and 'bucket') in Portuguese. Although he is able to supply the Portuguese words, JAM shows confusion when he talks about the names of the languages and asks his mother for clarification:

(118)
*JAM: <Portuguese["]>[@en] <é português["] o inglês["]>[@pt] ? [+ ep]
%add: MOT
"Portuguese" is "Portuguese" or "English" F043: L537

This uncertainty of how to refer to his two languages manifests itself throughout the particular file from which the example above is drawn and is also evident in another file (File 048) where JAM's mother is quizzing him about what he will be speaking to whom when he goes to visit England. The following excerpt perfectly illustrates how JAM appears to struggle when asked to think and talk about Portuguese and English metalinguistically. After having chatted with JAM about what he will be doing in England MOT then asks about the language he will need to use with his English cousins, Jake and Max:

(119)
*MOT: <and, <do you>[/] do you know what you're gonna[: going to] speak to Jake@pn and Max@pn>[@en]?
%add: JAM
*MOT: <are you gonna[: going to] speak Portuguese or English>[@en]?
%add: JAM
*JAM: <Portuguese["]>[@en] <é português["]>[@pt]? [+ ep]
%add: MOT
"Portuguese" is "Portuguese"?
*MOT: yeah[@en].
%add: JAM
*JAM: <English["][/] English["] is>[@en] <inglês["]>[@pt]? [+ ep]
%add: MOT
"English", "English" is "English"?
*MOT:uhhuh[@en].
%add: JAM
*JAM: <português>[@pt].
%add: MOT
Portuguese.
*MOT: 0 [=! sharp intake of breath, showing surprise].
*MOT: <you're gonna[: going to] speak Portuguese to them>[@en] +!?
%add: JAM
*JAM: yeah[@en], <só português>[@pt]. [+ ep]
%add: MOT
Yeah, just Portuguese. F048: L81-100

Although this time he gets his labels right in terms of translation equivalents, JAM still shows confusion as to what these labels actually mean. In the ensuing discourse (not shown here) JAM is able to identify that his cousins speak like 'Mummy speaks' (line 108) but is unable to label this language appropriately. In order to investigate this matter further I carried out two types of analyses: a frequency analysis to see how often JAM made references to both languages and a kwal analysis to see how they were actually used in his utterances. The frequency analyses of the four language words 'Portuguese', 'português', 'English' and 'inglês'[162] returned the following numbers of occurrences: 25 for 'Portuguese', 25 for 'português', 5 for 'English' and 40 for 'inglês'. By then examining these occurrences in the utterances[163] it was then possible to spot patterns as to how and why JAM used these words. Although it would be interesting

[162] Obtained by the command, freq @ +t*JAM +s"Portuguese" +u, where 'Portuguese' was then substituted by each of the other three words in turn ('português', 'English' and 'inglês').
[163] Obtained by the command, kwal @ +t*JAM +s"Portuguese" +u +d1, where 'Portuguese' was then substituted by each of the other three words in turn. By adding the number 1 to +d, the line number for each utterance was included in the output faciliating its location in the original file.

to report in detail on the patterns found and illustrate them with examples, here there is only space to summarise my observations.

In the first 42 files of the corpus language references are few: out of a total of 13 tokens, the Portuguese labels account for 11 of the occurrences ('inglês' occurring 7 times and 'português' 4 times) while the term 'English' is used twice. It is in Files 043 and 048 that we find the highest concentration of language references with over a third (35) of all the references (95). The majority of these tokens are for the word 'inglês' (19 occurrences), 'português' appearing in second place with 9 tokens and the English labels accounting for relatively few tokens ('Portuguese' occurring 5 times and 'English' occurring 2 times). From the example and excerpt discussed above it is not suprising that there are so many language references in Files 043 and 048 given the metalinguistic nature of these two conversations between JAM and his mother. However, it is significant to note that in the first 48 files JAM uses mostly Portuguese labels (39) as opposed to English labels (11). Taking into account this quantitative data and from the qualitative evidence shown above (in the example utterances) it appears that over this period of time JAM is more confident in his use of the Portuguese labels. It is exactly when he is obliged to use the English labels for the languages that he becomes confused.

This confusion, however, appears to be temporary, as an analysis of the remaining occurrences revealed. In fact, with no further recorded conversations between mother and son specifically about language issues, the language labels are seen to occur much less and are used mostly pragmatically. JAM's use of 'Portuguese', the source of earlier confusion, is given a meaningful boost while on holiday in England: while talking to his father over the telephone (Files 062, 069, 071, 073 and 074) JAM has the need to turn to his mother and make requests of the 'How do I say.... in Portuguese' type. These types of requests account for 16 of the occurrences of 'Portuguese' and show how much influence immersion in a different language context can have on such usage. However, back in Brazil and seven months later English has clearly lost ground to Portuguese as evidenced in a short telephone conversation between JAM and his English Grandmother. This time, having difficulty recalling certain English words, JAM repeatedly turns to his mother and asks for help using the Portuguese request 'Como é que é ... em inglês?' ('How/What is ... in English?'). He does this a total of 8 times!

There is no evidence in the output for the files after number 048 to suggest that JAM still has difficulty in referring metalinguistically to either language. This is likely to be a reflection of both developing linguistic ability and increasing metalinguistic language awareness. However, when we look at the last occurrence of the word 'Portuguese' we see a new language reference being used by JAM. The following utterance occurs during a telephone call between JAM and his Brazilian grandfather approximately five months are having moved to England. Being asked about the name of his best school friend (Joshua), JAM says the following:

(120)
*JAM: <eu não sei o>[@pt] Por(tuguese)[@en][//] <em brasil(eiro)[/] brasileiro>[@pt]. [+ pep]
%add: VOV
I don't know the Por(tuguese), in Brazil(ian) Brazilian. F118: L67

He begins to say that he does not know how to say it in 'Portuguese' but retraces and then uses the term 'em brasileiro'. Although the English equivalent of this language label, 'Brazilian' may be used erroneously by those assuming that in Brazil one speaks 'Brazilian', it would be very strange, and incorrect, for a Brazilian native speaker of Portuguese to refer to his language as 'brasileiro'. In the data available there is no evidence that JAM had done so when living in Brazil[164]. A plausible explanation for this occurrence would be that JAM's school peers, and even some adults, had made reference to JAM's other language as being 'Brazilian', perhaps when asking him how to say things 'in Brazilian'. It seems that such usage may be influencing JAM, causing him to transfer the English term into Portuguese.

As seen above, by carrying out a simple analysis of JAM's references to languages, we have been afforded insights into how complex it can be for a bilingual child to make sense of his languages metalinguistically. While he might show competence when it comes to actually using both languages (whether in monolingual or bilingual mode) it appears that *talking* about this usage is another matter, as revealed by some of the examples shown in this section.

Again, it is important to mention the value of the methodology used in this study in allowing such observations to be made. While the use of the metalinguistic coding in the corpus and the CLAN software enabled an efficient and effective

[164] The following two analyses, kwal @ +t*JAM +s"brasileiro" + u +d' and kwal @ +t*JAM +s"Brazilian", revealed no other occurrences of 'brasileiro' in the corpus and only two occurrences of 'Brazilian', used to refer to people's nationality.

analysis of how JAM used the device of quoting in his CS utterances, it also flagged up the issue of his metalinguistic awareness, provoking the subsequent longitudinal analysis of the language labels he uses. Through the use of specific command lines, the output data (in the form of frequency lists or utterances) is immediately available for analysis and any need to locate a particular utterance in its context is achieved by quick access to the original file from within the CLAN window.

The need to refer back to the original files in order to understand the use of a metalinguistic code was more frequent when it came to the output provided for MEG as will be seen in the following section.

6.5.2 Metalinguistic usage in MEG's code-switched utterances

With 67 occurrences of the **["]** code in MEG's CS data, it is clear that she made significantly more use of this linguistic device than her brother. Of this total number of occurrences, 37 of the codes (55%) were related to quoting while the remaining 30 (45%) were related to metalinguistic usage. In terms of proportions, this usage was similar to JAM's, the percentages for whom were 62% for the former and 38% for the latter.

Looking first at MEG's use of quoting in CS utterances, we find cases of her switching to Portuguese in order to quote her Brazilian school teachers. In this particular example below MEG is explaining to her mother how the teacher has told them to perform a dance:

(121)
*MEG: <yes, we have to put <like this>[//] arms like this and>[@en] <<vai girando, girando, sem parar>["]>[@pt] . [+ ep]
%add: MOT
Yes, we have to put like this, arms like this and "go spinning, spinning around without stopping". F010: L197

It is clearly easier for MEG to revert to Portuguese when talking about her Brazilian school-based events rather than attempt to translate her teacher's words into English. Two years later when MEG is telling her father over the telephone about her *English* school-based events, we again see how the language of the school environment can encourage code-switching with her bilingual interlocutor. In this case MEG is telling her father about the sorts of things they can take in their packed lunch and refers to the 'healthy eating' campaign which her school promotes:

(122)
*MEG: <não, na minha escola tem que ser umas coisas assim de>[@pt] <<healthy eating>["]>[@en] . [+ pe]
%add: PAI
No, in my school it has to be some things like "healthy eating". F102: L429

Of course, MEG only has the option to code-switch above because both parents are bilingual. It is difficult to say how MEG would have expressed the above monolingually, but it would clearly have been more challenging and less economic, linguistically.

Even where the translation of the quoted word or words would not have represented a challenge for MEG, there are times when such translation is undesirable. An example of this is found in the CS utterance below where, following a telephone conversation with her father, MEG is reporting back to her mother about how her guinea pig starts squeaking when her father calls out 'cachorro'.

(123)
*MEG: <and as well <when he>[//] he said when he comes in he says>[@en] <cachorro["]>[@pt] <and he says she starts to <mi+mi+mi>["]>[@en] . [+ epe]
%add: MOT
And as well when he, he said when he comes in he says 'dog' and he says she starts to 'mi mi mi'. F060: L184

From the gloss we can see that 'cachorro' actually means 'dog'. Contextual information (provided in the file itself) tells us that this Portuguese word (said with a Spanish accent) had been adopted by the father as a pet name for MEG's guinea pig (called Biju) and here MEG is simply quoting how her father calls out to the animal, and how the latter responds, with 'mi mi mi'. Only by having access to extra-linguistic information do we understand why MEG does not translate this quoted word into English.

It is perhaps rather surprising to find that a few of MEG's CS utterances actually reveal the use of quoted Spanish words and phrases. The occurrence of Spanish words in both siblings' CS utterances was first brought to light in the frequency analyses of the CS codes where the 's' for Spanish appeared in the postcodes (see Tables 15, 16 and 17 in section 5.2.1). Then, when carrying out the frequency analyses of the metalinguistic code (see Table 26 in section 5.2.5.2) we saw that some of these Spanish words appeared in these word lists. I purposely did

not discuss JAM's use of these Spanish words in the previous discussion as it makes sense to discuss the siblings' usage together, as will be seen below.

A simple frequency analysis of the Spanish words (coded with **@sp**)[165]occurring in the corpus revealed a total of 41 tokens and 8 word types (the number of occurrences for each type is in brackets): 'burro' (14), 'mucho' (8), 'hay' (5), 'mantequilla' (5), 'dos' (3), 'que' (3), 'tres' (2) and 'Manuel@pn' (1)[166]. A KWAL analysis[167] then showed that all of these 38 tokens actually came from one single file (079) and were distributed over 22 utterances, of which 11 were uttered by MEG, 10 by JAM and one by PAI. After consulting the original file what comes to light is that MEG and JAM are quoting from a British sitcom called Fawlty Towers which features a Spanish waiter, Manuel, who works at a hotel owned by a Basil Fawlty. It is MEG who instigates this particular dialogue at breakfast time when she comments that her mother has put a lot of butter on her bread. Drawing on what Basil had said to Manuel about there being too much butter on the breakfast trays he was taking to the guests, MEG says the following:

(124)
*MEG: <<hay mucho burro>["]>[@sp] <on my bread>[@en] . [+ se]
%add: MOT
'*There is a lot of donkey' on my bread.* F079: L610

From the ensuing conversation it becomes evident that when she says the above MEG erroneously thinks that 'burro' means 'butter' (as did Basil). Prompted by this quote, JAM then joins in with MEG as they both begin recalling this particular scene from the sitcom and start quoting the exchange that occurred between Basil and Manuel (using monolingual English quotes as well as bilingual ones). JAM becomes very involved in quoting the original exchange word for word and this gives rise to the following utterance addressed to his father:

(125)
*JAM: +" <but a>[@en] <burro["]>[@sp] <is a eeyore>[@en] . [+ ese]
%add: PAI
"*But a 'donkey' is a 'eeyore'".* F079: L714

[165] freq @ +s"<@sp>" +u +o
[166] Eleven quote codes (["]) also appeared in the word list and were manually excluded from the original token total (52).
[167] kwal @ +s"[@sp]" +u +t%add

It is after this quote that we see a divergence in focus between the siblings. While JAM is happy to continue quoting Manuel's 'funny' words to his father, MEG has clearly began to reflect on what these quotes mean and why they are cause for amusement, as we can see in the following two questions she poses to her mother:

(126)
*MEG: <why did he say that Manuel@pn>[@en] <<but a>[@en] burro[@sp] <is a eeyore>[@en]>["] . [+ ese]
%add: MOT
Why did he say that Manuel "but a 'donkey' is a 'eeyore'"? F079: L737

(127)
*MEG: <but what is>[@en] <mantequilla["]>[@sp] ? [+ es]
%add: MOT
But what is 'butter'? F079: L783

After her father explains that 'burro' is Spanish for 'donkey', MEG then shows her understanding of the humour underlying the original exchange between Basil and Manuel by applying this humour to her own utterance:

(128)
*MEG:<têm muitos jumentos no meu pão>[@pt].
%add: PAI MOT
There are a lot of donkeys on my bread. F079: L755

And then when she learns from her mother that 'mantequilla' is the Spanish word for 'butter' she is then able to reformulate her initial utterance (which instigated this whole dialogue) using 'mantequilla' instead of 'burro' to say that her mother had put too much butter on her bread:

(129)
*MEG: <there's too much>[@en] <mantequilla["]>[@sp] <on my bread>[@en] . [+ ese]
%add: MOT
There's too much 'butter' on my bread. F079: L791

Such creative use and application of metalinguistic understanding is not evident in JAM's utterances: it appears his aim is simply to amuse his father by quoting from a sketch which he had found funny because of the way Manuel speaks and because of how the latter is treated by Basil.

The main aim of the investigation above has been to reveal why Spanish words should appear in the metalinguistic frequency analyses of CS utterances. Restricted to one particular conversation we have learnt that both siblings are quoting

lines from a television show. However, the detailed study of the dialogue has also served to highlight a difference between the siblings that is evident elsewhere: that MEG's linguistic awareness is more developed than her brother's, shown above by her considered reflection of the meaning of the Spanish words she is quoting and her incorporation of this understanding into her own productive output.

A further difference between the siblings is shown in terms of the types of quoting we can see in the output for MEG from the KWAL analyses. Whereas all of the metalinguistic codes which referred to quoting in JAM's data (15 occurrences) related to quotes from speakers, in MEG's CS data we find her quoting from other sources. While 24 of her total of 37 codes refer to speakers, the remaining 13 refer to quotes of written words. Amongst these are names of books, television programmes, games and products and their quoted use is more often than not the source of a code-switch. Although we have seen that MEG will self monitor to ensure that she does not code-switch unnecessarily, the case of quoting written words seems to serve as an exception. This can be seen in the following utterance where MEG is recounting to her father over the telephone about a flyer that had been posted through the door about a lost cat:

(130)
*MEG: <e[/] e a gente, pelo>[@pt] post[@en] <veio um papelzinho dizendo>[@pt] <<lost cat, name, very gentle, tortoise+shell, black and white, name say>["]>[@en] . [+ pe]
%add: PAI
And, and we, through the post came a bit of paper saying "lost cat, name, very gentle, tortoise-shell, black and white, name say". F104: L305

The translation of such English words into Portuguese would not have represented a challenge to MEG (apart from possibly 'tortoise+shell') but she clearly prefers to stick to the original. The quoting of the original name of a card game is the source of a very rare code-switch with a monolingual speaker, as can be seen in this example where MEG is talking to her Brazilian cousin Sara:

(131)
*MEG: <a Mamãe@m disse que vamos brincar de>[@pt] <donkey["]>[@en] . [+ pe]
%add: SAR
Mummy said that we are going to play 'donkey'. F047: L20

Looking at the dialogue from where this particular utterance was drawn (File 047), we see that MEG, JAM, MOT and SAR are playing the card game 'Donkey'. With the

pack of cards having come from England (and having been used at least once previously by the family[168]) MEG sticks to the original English name although it is clear from later on in the conversation that Sara already knows that 'donkey' means 'burro': she uses 'burro' 3 times throughout the game. Thus there was really no need for MEG to effect the translation of 'donkey' for her cousin.

Another example of MEG quoting from the written word can be seen in the following CS utterance where she tells her father that she had seen a sign which had 'guinea pigs' written on it:

(132)
*MEG: <e aí olhei para esse[/] esse[//] essa placa bem grandona e aí[/] e aí eu diz lá>[@pt] <<guinea+pigs>["]>[@en], <<ai meu deus eu vou ler a coisa toda>["]>[@pt] . [+ pep]
%add: PAI
And then I looked at this, this, this really big sign and then, and I said there 'guinea pigs', 'Oh my God I'm going to read the whole thing'. F106: 358

As can be seen from the three examples shown above and from all the other examples discussed so far, the quoting of both spoken and written words (37 occurrences) accounts for much of the code-switching in which MEG engages. However, another significant source of code-switching, revealed by an examination of the CS utterances containing the remaining 27 metalinguistic codes, is that of direct metalinguistic usage, as will be shown below.

Ten of the remaining codes are found in the following type of utterance where MEG is making a metalinguistic request, that is, asking how to say something in either English or Portuguese.

(133)
*MEG: <how do I say>[@en] <coração["]>[@pt] <in English>[@en][=! whispers] ? [+ epe]
%add: MOT
How do I say 'heart' in English? F024: L141

All of the 10 requests are for single words, the words themselves being the only code-switched word in an otherwise monolingual utterance. What is interesting about the particular example above is the fact the MEG whispers her request. In this file (024) we find two more requests for English words, both whispered. It appears that MEG does not wish to be recorded making such requests. On consulting the original file we learn from MOT's second utterance that the purpose of the recording was for

[168] The command line kwal @ +s"donkey" +u +t%add revealed that this card game had been played in File 027, the participants being MEG, JAM and MOT.

MEG to recount a dream she had had. Although MEG is addressing her mother, she is clearly aware of the recorder and therefore of a potential audience. Her self-monitoring becomes acute - she clearly wishes to avoid switching to Portuguese while telling her dream. It is likely that she believes that by whispering her request in an aside, this use of Portuguese, and the fact that she has to ask for their translations, will not come out in the recording.

MEG's awareness of which words come from which language and her ability to apply the appropriate language labels are shown very early on in the corpus data, as the following CS utterance reveals:

(134)
*MEG: because[@en] <computador["]>[@pt] <is Portuguese and computer["] is English>[@en] . [+ epe]
%add: MOT
Because 'computer' is Portuguese and 'computer' is English. F005: L143

In response to her MOT's question, MEG is 'explaining' how she knows which words comes from Portuguese and which come from English. This is after having played a language label game where MEG had to tell her mother which language words came from. A look at this utterance in its context shows that MEG appears to judge words by their sounds and associates them with the people who might say them. For example, she knows that the word 'cadeira' is Portuguese because Inês (the Brazilian maid) uses it and she is from Brazil[169]:

It is in conversations such as these between mother and daughter that we find many of the remaining 17 occurences of metalingustic codes. This is not surprising given the metalinguistic nature of these informal chats about language. This would also explain the frequency with which MEG uses the English language labels as opposed to their Portuguese equivalents. A frequency analysis of the four language labels (the same that was carried out on JAM's utterances)[170], revealed the following number of occurrences for each label: 51 for 'Portuguese', 3 for 'português', 58 for 'English' and only 4 for 'inglês'. These frequencies contrast quite dramatically with those found for JAM (25, 25, 5 and 40) and a KWAL analysis of these occurrences[171] confirmed that only once did MEG use a Portuguese language label when

[169] *MEG: <<because I>[//] (be)cause>[@en] Inês[@pt] <she is[/] is[///] lives in Brazil (a)n(d) Brazil's Portuguese>[@en] (F005: L143)
[170] See footnote 162.
[171] See footnote 163.

addressing her mother. The other 6 uses of Portuguese language labels were addressed to her father (4 occurrences) and her Brazilian Grandfather (3 occurrences). The CS utterance below shows that even when MEG did refer to 'inglês' on a single occasion with her mother, she retraces and uses 'English':

(135)
*MEG: <you said in>[@en] <inglês>[@pt][//] <in English, okay["]>[@en] . [+ epe]
%add: MOT
You said in English in English 'okay'. F078: L614

Examining this utterance in context we see that MEG seems to be reproaching her mother for the latter's use of 'okay' in English to the Brazilian maid (ARL). As far as MEG is concerned this was not the correct language to use.

Such examples coupled with the quantitative frequency data have served to show that again we are seeing the consistency with which MEG uses English with her mother, even in terms of talking about language use. Unlike the data for JAM, which revealed his struggle to get to grips with both his languages conceptually, the data for MEG reflects a comparatively heightened language awareness which is evident over the whole time span of the corpus.

In this section, the analysis and subsequent discussion of the occurrences of the metalinguistic code ["] in the siblings' CS data has proved to be very productive. We have seen that for both JAM and MEG an important function of code-switching is that of being able to quote a speaker's original words. In MEG's case this use is also extended to include the quoting of written words. However, it is perhaps the analysis of those CS utterances where the code is marking metalinguistic usage that has provided us with the greatest insights into the differences in how the children use and understand their two languages. Combined with the additional analyses carried out on the use of the language labels, we have learnt from the data that MEG appears to have conceptually compartmentalized her two languages from early on and has no difficulty reflecting on this compartmentalization (based of corpus data before the age of 6). Although it is understandable that JAM, being two and a half years younger, has yet to develop the same level of language awareness as his sister, the data does suggest that there are factors which may exert a strong influence on this development: through the longitudinal analysis of the language labels we saw the

effect of linguistic environment on the way JAM referred to his languages, even when addressing the same speaker.

Before leaving this section I will briefly look at the data for MOT which showed a potentially significant relationship between metalinguistic referencing and code-switching.

6.5.3 Metalinguistic usage in MOT's code-switched utterances

The original frequency analysis of the metalinguistic codes in MOT's utterances had revealed that 12% (57) of all of her ["] codes were seen to occur in CS utterances. With only 1.8% of all MOT's tokens occurring in CS utterances, this means that such metalinguistic usage appears to be a particular feature of her code-switching. As for JAM and MEG, KWAL was used to output all of MOT's CS utterances containing the metalinguistic code[172]. The resulting 51 utterances were then analysed in terms of addressee and the function of the coded elements. Apart from a single utterance addressed to PAI, all of the remaining 50 utterances were addressed to JAM (29) and MEG (21). An analysis of the 29 utterances addressed to JAM revealed that on 8 occasions the ["] code was being used to mark elements in questions of the following type:

(136)
*MOT: <how do you say>[@en] <<me dá>["]>[@pt] ? [+ ep]
%add: JAM
How do you say 'give me'? F048: L411

The function of these questions was to encourage JAM to say, in English, something he had originally said in Portuguese. This contrasts with the type of metalinguistic questions MOT asks her daughter, as seen below:

(137)
*MOT: <I mean, why does>[@en] <cadeira["]>[@pt] <sound Portuguese>[@en] ? [+ epe]
%add: MEG
I mean, why does 'chair' sound Portuguese? F005: L102

It is in MEG's answer to this question that we learn how she associates words phonetically with speakers (see discussion above and footnote 169 for her answer). Such metalinguistic probing is not evident in MOT's utterances addressed to JAM.

[172] Kwal @ +t%MOT +u +s"[+ *]" +s'["]' +t%add +fMOTmeta

Another noticable difference is in the number of times the ["] code is seen to mark the answer to a 'How do you say ...?' request, such as in the following example:

(138)
*MOT: <I think it's>[@en] <trem+a+vapor["]>[@pt] . [+ ep]
%add: JAM
I think it's 'steam train'. F071: L343

There are 10 such uses in the data for JAM as addressee but only one where MEG is the addressee. These results tally with the earlier finding in this section that JAM, but not MEG, made frequent metalinguistic requests to his mother while he was engaged in conversation with his father over the telephone.

In the remaining CS utterances which contain the ["] code, we see that the other major function is that of quoting another's speech. However, unlike JAM and MEG who were seen to quote from a wide variety of sources (both spoken and written in MEG's case), the corpus data only ever shows MOT quoting her children, usually for clarification purposes, as illustrated by the two examples shown below:

(139)
*MOT: <what was>[@en] <<cem libras>["]>[@pt] ? [+ ep]
%add: JAM
What was 'a hundred pounds'? F119: L227

(140)
*MOT: <oh tv["], I thought you said>[@en] <dever["]>[@pt] . [+ ep]
%add: MEG
Oh 'TV', I thought you said 'homework'. F083: L145

Such quoting can be found in 9 of MOT's CS utterances addressed to JAM (29 in total) and 10 addressed to MEG (21 in total).

In the discussion of the metalinguistic word frequency data (see 5.2.5.2), it had already been established that 25% of all of the Portuguese words used by MOT in CS utterances were coded with ["]. In this section we have examined MOT's utterance data in detail and it is now possible to see how much of her code-switching activity into Portuguese is actually triggered by her children, whether through her quoting of their words or referring to their language metalinguistically in the form of questions and answers.

For the sake of comparison, a frequency analysis of the four language labels was also performed on all of MOT's utterances[173] and resulted in the following frequencies: 72 occurrences of 'Portuguese', 4 of 'português', 147 of 'English' and only 1 of 'inglês'. Such relative frequencies are similar to those found for MEG (51, 3, 58 and 4) but contrast markedly with those found for JAM (25, 25, 5, 40). This is perhaps further evidence to suggest that MEG is much more linguistically in tune with her mother than her brother is - she makes reference to her two languages in a similar fashion to her mother.

The analysis of the metalinguistic code in this section has been extensive and detailed and as such has made an enriching contribution to the current investigation of code-switching in the corpus. It has also allowed for the detection of subtle differences in how the siblings use and understand their two languages metalinguistically. Although the evidence from the longitudinal analysis of the language labels supports the idea that these differences are related to linguistic maturation, we have also seen the role contextual factors, such as the linguistic environment, can have in affecting the development of language awareness.

As has become evident from the discussions in this chapter, it is only through an utterance-level analysis of the siblings' codeswitching behaviour that we can begin to learn more about the structural nature of their codeswitches and uncover the motivations behind their use of this bilingual phenomenon. By examining the utterances in their wider linguistic context I have been able to consider the extent to which each of the siblings' code-switching behaviour is affected by the local discourse environment as well as by the wider sociocultural environment. And by incorporating a longitudinal perspective into my analysis of some of JAM and MEG's codeswitches I have been able to explore how developmental aspects may affect the outcome of their bilingual utterances. Although throughout this dissertation I have consistently highlighted the importance of taking into account the addressee variable when examining CS data, it is also important to consider how the language practices of the addressees themselves might influence the code-switching behaviour of the speakers under analysis. In the case of my study this involves examining the language behaviour of MOT and PAI, both when interacting with their children and

[173] Using the following basic command line, each language label was then substituted in: freq @ +t*MOT +s"Portuguese" +u

when interacting with each other. To this end, in the following, penultimate, chapter I offer an analysis of the parents' code-switched data. I also offer an utterance-level examination of the code-switching occurring between the siblings in order to search for explanations for the earlier quantitative findings which were discussed in Chapters 4 and 5.

7. Analyses of the parents' code-switching and of that occuring between the siblings

The focus of the utterance-level analyses presented in Chapter 6 was on the code-switched utterances of the two main informants of this study, JAM and MEG. While some of the analyses were addressee specific and looked particularly at the siblings' CS utterances addressed to MOT and PAI (6.1. and 6.2), for other analyses the variable of time was considered to be a more relevant factor (6.3, 6.4 and 6.5). In both cases it was the siblings' data that was under scrutiny although a brief analysis of the occurrence of the [“] symbol in MOT's code-switched utterances was also included (6.5.3). In the current chapter, I will examine the code-switching practices of MOT and PAI, both when addressing their children and when addressing each other. I will also look at the nature of the code-switching occurring between JAM and MEG in order to see whether an utterance-level analysis will support what the various quantitative results so far suggest - that neither English nor Portuguese can be said to be taking on the role of the Matrix Language in their bilingual interactions. Thus, the speaker/addressee combinations focussed on in the following sections are the following: MOT/JAM, MOT/MEG, JAM/MEG, MEG/JAM, MOT/PAI, PAI/MOT, PAI/JAM and PAI/MEG. As the output for the first two combinations was greatest I will begin with the results pertaining to MOT's code-switching with her children.

7.1 MOT's code-switching with her children

Frequency anayses had shown that MOT engaged in relatively little code-switching with her children (see 4.1.4), addressing them almost exclusively in English throughout the time period of the study. In this section I will briefly examine those utterances where code-switching did occur in order to determine the nature and motivation behind such (limited) usage. First of all I asked KWAL to provide me with two sets of CS utterances - those addressed to JAM and those addressed to MEG[174]. Although the output returned 108 CS utterances for JAM and 55 for MEG, before analysing these utterances qualitatively I decided to manually exclude any utterances which were addressed to both JAM and MEG, or indeed to any other additional interlocutor. By doing this I could potentially eliminate the influence of the

[174] kwal @ +t%add +t*MOT +s"JAM" +s"[+ *]" +u +fmot and Kwal @ +t%add +t*MOT +s"MEG" +s"[+ *]" +u +fmot

presence of other interlocutors on MOT's code-switching, giving, therefore, a more accurate picture of her CS practices with her children[175]. This exclusion resulted in a reduced total of 86 utterances with JAM as single addressee and 39 with MEG as single addressee.

On an initial examination of the data (in fact, after looking at the first three CS utterances addressed to JAM) I realised that it would be useful to exclude a futher set of utterances before proceeding with my qualitative analysis. These particular utterances involved the MOT's use of 'olha' ('look'), most often reduced to 'o(lha)' (i.e. the initial open vowel sound). In the majority of cases, this Portuguese discourse item represented the only contribution to an otherwise monolingual English utterance and therefore I decided to exclude these CS utterances from the totals. Of MOT's 86 utterances addressed to JAM, a total of 39 were found to contain 'olha' as the only Portuguese word. Of the CS utterances addressed to MEG, only 8 were found to be of this nature. This relative difference in use of 'olha' may simply reflect the MOT's need to make more use of such an attention-directing device when talking to her younger son. What is of greater interest here is what the remaining utterances (47 for JAM and 31 for MEG) reveal about her more productive use of code-switching with her children.

7.1.1 MOT's code-switching with her son

An analysis of MOT's 47 CS utterances addressed to JAM revealed that her use of Portuguese was mostly restricted to single word insertions, a typical characteristic of an Embedded Language. And out of the total of 69 Portuguese words used by MOT, 48 were actually coded with ["], showing her use of Portuguese for the purpose of quoting words/phases or for metalinguistic references. Two examples are shown below:

(141)
*MOT: <why[/] why do you say>[@en] <cinco["]>[@pt] ? [+ ep]
%add: JAM
Why, why do you say "five"? F018: L132

(142)
*MOT: <how do you say err>[@en], <<me dá o carro aí>["]>[@pt] ? [+ ep]
%add: JAM

[175] See 8.5 for more discussion on this exclusion.

How do you say err "Give me that car"? F048: L399

In the first example, MOT asks JAM why, after counting to four in English, he then uses 'cinco' instead of 'five'. The second example shows one of only four CS utterances where MOT quotes entire phrases (accounting for 13 of the 48 Portuguese words coded with ["]). The remaining 21 Portuguese words used by MOT when code-switching with JAM are mostly nouns intrinsically linked to their shared sociolinguistic and cultural environment, as illustrated by the two examples below:

(143)
*MOT: <just get them ready for the party and then I'll go and get some>[@en] rapadura[@pt] . [+ ep]
%add: JAM
Just get them ready for the party and then I'll go and get some sugarcane fudge. F011: L67

(144)
*MOT: <and all of the sch(ool), all of your friends in your class and the ones from>[@en] <alfabetização>[@pt] . [+ ep]
%add: JAM
And all of the school, all of your friends in your class and the ones from the learning to read and write class. F040: L56

In the first example MOT refers to a Brazilian sugarcane sweet for which there is no satisfactory English translation and in the second example she refers to the first year of Brazilian primary education where children are taught to read and write[176]. The total absence of grammatical items in this group of 21 Portuguese words reinforces how classic a role Portuguese is performing in MOT's CS utterances addressed to her son.

7.1.2 MOT's code-switching with her daughter

On analysis of the 31 CS utterances MOT addresses to her daughter, we again find classic use of code-switching. Of only 44 Portuguese tokens, there are 34 coded with ["], 13 of which appear in two utterances. One of the latter is shown below:

(145)
*MOT: <so how do you say>[@en] <<cabelo mais escuro de que a Sara@pn>["]>[@pt] ?

[176] Although in the British Primary school system, there is a term for the equivalent class ('Reception'), MOT would only have become familiar with this term when she returned to England and placed JAM and MEG in their English school.

[+ ep]
%add: MEG
So how do you say "darker hair than Sara"? F038: L444

While in the above utterance MOT is prompting MEG to try and express a Portuguese phrase in English, in the example below she is pointing out that the correct agreement on the Portuguese word for 'thank you' should be 'obrigada' (with the feminine 'a' ending as opposed to the masculine 'o' ending).

(146)
*MOT: <isn't it>[@en] <obrigada["]>[@pt] ? [+ ep]
%add: MEG
Isn't it thank you? F049: L160

Referring back to the original transcript we learn that MOT is actually questioning MEG's quoting of her English Grandmother who had told her that on the plane to Brazil she had said 'Obrigado' to the flight attendent. This makes more sense as one would not expect MEG to use the masculine form when expressing her own gratitude.

Most of MOT's use of Portuguese with her daughter is metalinguistic: only 10 futher tokens are not coded by ["]. Again, as for the CS utterances addressed to JAM, most of these 10 tokens are socially or culturally bound, for example, 'tatu' (a Brazilian armadillo), 'Nescau' (a branded chocolate drinking powder), 'flocos' (a chocolate chip ice-cream) and 'aventureiros' (a television programme).

From the qualitative analysis of MOT's CS utterances addressed to her children, it has been possible to ascertain that her code-switching practice is very restricted, both in terms of quantity and variety. English very clearly takes on the role of the Matrix Language with Portuguese being used very sparingly and for specific purposes. This finding supports what was revealed by the quantitative analyses discussed in Chapters 4 and 5.

7.2 Code-switching between the siblings

In the same way that any multi-addressed CS utterances were removed from the above qualitative analysis, when examining the data for the siblings I decided to focus on only those utterances which had a single addressee. Therefore, when looking at JAM's CS utterances to MEG, if any other speaker codes were found on the %add tier, these particular utterances were excluded from the analysis. Similarly,

only MEG's CS utterances addressed solely to JAM were the focus of the analysis reported on in this section.

7.2.1 JAM's code-switching with his sister

Of the 48 CS utterances found in the output from the KWAL analysis[177] for JAM, there were 27 addressed exclusively to MEG. Eleven of these consisted of the insertion of single English words into otherwise Portuguese utterances. These eleven words were typical of an Embedded Language contribution: Monday, not, press, beyblade[178] (x2), hole, here, burnt, buy, invisible and flowers. A further four CS utterances involved retracings where JAM switches in order to provide the translation equivalent: in three of the cases JAM switches from English to Portuguese (look to olha, were I to estava and wait to espera; in the fourth case the switch is in the opposite direction (tá bom to okay). Leaving aside this fourth switch, we so far have evidence (14 utterances) to suggest that when code-switching with his sister Portuguese plays a more dominant role, that of the Matrix Language.

Even in the following utterance where English is seen to contribute more tokens (3) than Portuguese (2), the lack of the auxilary 'is' before 'going' and the maintenance of the Portuguese syntax in 'me balançar' implies that it is Portuguese which is exerting more influence in this utterance:

(147)
*JAM: <Mummy@m going to>[@en] <me balançar>[@pt] . [+ ep]
%add: MEG
Mummy going to rock me. F045: L455

Contextual factors are seen to have an impact on the dominance of Portuguese in JAM's code-switching with his sister, as shown by the following example where JAM is playing 'Guess who?' with MEG while on holiday in England.

(148)
*JAM: <is the>[@en] <cabelo marrom>[@pt] ? [+ ep]
%add: MEG
Is the hair brown? F053: L1753

The fact that English is supplying two grammatical elements and Portuguese is contributing with a noun and adjective suggests that here the former is acting as the

[177] kwal @ +t*JAM +t%add +s"MEG" +u +d +s"[+ *]" +fJAM
[178] A spinning top toy.

Matrix Language and the latter is an Embedded Language island. However, I would like to argue that in this case, Portuguese is still playing the more dominant role in this utterance. An examination of this particular transcript (File 053) reveals that MEG had been playing 'Guess who?' with her English aunt before she begins playing with her brother and appears prepared to carry on asking the yes/no questions in English. However, JAM has not been 'primed' in the same way and his first question is automatically in Portuguese (line 1334). Prompted with suggestions for questions from his mother, JAM does then ask 3 questions in English (lines 1425, 1442 and 1596) but they contain errors, for example 'JAM: is he got blue eyes?'. While MEG continues asking questions in English, JAM then uses two CS questions: 'JAM: is it preto?' (black) and the one shown above. Although his use of 'is' at the beginning of these two CS questions is felicitous, its use in the monolingual English question mentioned above implies that perhaps JAM is using 'is' as a generic way of starting his yes/no questions. As the game continues MEG begins to get frustrated at how long JAM takes to formulate a question and respond to her questions. To speed things up she decides to switch to Portuguese, reformulating and translating her original English question ('Is it a boy?') into Portuguese ('é um menino?'). From then on both JAM and MEG's questions to each other are in monolingual Portuguese.

If we consider that this game takes place two days after the siblings' arrival in England, it is not surprising that JAM appears to struggle to formulate appropriate questions in English. Although his use of 'is' in the CS questions shows a willingness to comply with the language expectations of the situation, it does not really lend support to the idea that English is now playing the role of the Matrix Language in his CS interactions with his sister. The fact that both siblings revert to monolingual Portuguese at MEG's signal is further evidence that Portuguese, and not English, still plays the more dominant role.

In four out of the 27 CS utterances JAM addresses exclusively to MEG, we find two occurrences of the formulaic expression 'let me see' and two cases of the adverbial phrase 'very well', one example of which is shown below:

(149)
*JAM: <não está prestando>[@pt] <very well>[@en], não[@pt] . [+ pep]
%add: MEG
It's not working very well, no. F058:L305

Here, JAM inserts the adverbial phrase into an otherwise Portuguese utterance. It is likely that two weeks of immersion in English has led him to use this expression instead of the Portuguese equivalent 'muito bem', its insertion facilitated by the word order equivalence existing across the two languages for this expression.

Although the immediate linguistic context can clearly affect the normal patterns of a speaker's language use, as evidenced above, it is possible to see just how longlasting this effect can be in the following CS utterance which was recorded almost two months after JAM's return to Brazil (from holidaying in England):

(150)
*JAM: <Meggie@pn # remember the seal>[@en] <que>[@pt] burped[@en]
[=! makes noise of seal burping] ? [+ epe]
%add: MEG
Meggie, remember the seal that burped? F084: L564

In a conversation at the dinner table, JAM recalls the time they went to a wild life park in England and saw a seal performing in a show. This had obviously been one of the more memorable and amusing experiences of their holiday and at the time there must have been plentiful discussion (with their English relatives and friends) about this seal that had 'burped'. It is not very surprising, therefore, that JAM should recall the incident in English, especially as he had never had the opportunity to see a seal in Brazil[179]. It is puzzling, however, why JAM should then choose to insert the Portuguese generic grammatical relative pronoun 'que', a typical contribution from the Matrix Language. Such usage could imply that Portuguese is still more dominant in this situation and that the four English words actually represent a kind of reported speech, primed by previous recountings of the occasion. This interpretation is further supported if one considers that JAM might be avoiding using the English relative pronoun 'which', which he had previously equated (erroneously) with 'que' (see section 6.3.2). Such avoidance implies JAM is still more grammatically confident in Portuguese.

Unlike most of the 27 CS utterances that JAM addresses to MEG, where Portuguese is easily identifiable as the Matrix Language, the example discussed above clearly represents more of a challenge for interpretation. However, by taking

[179] KWAL searches for JAM's use of the Portuguese equivalents of 'seal' ('foca') and 'burp' ('arroto') in the corpus (kwal @ +t*JAM +u +d +s"foca" and kwal @ +t*JAM +u +d +s"arrot*") returned zero occurrences of 'foca' but 4 occurrences of 'arroto(s)' and 1 of 'arrotou'. Although the absence of 'foca' in the corpus does not automatically mean it was not part of his productive vocabulary, the presence of the Portuguese equivalent of 'burp' confirms JAM's productive knowledge of this word.

into account contextual information, it is possible to shed more light on this apparent contradiction to his normal pattern of code-switching with MEG.

Explanations for the final three examples are equally reliant on access to contextual information. Occuring after the siblings' permanent move to England, the nature of these CS utterances reflects a change in the asymmetrical language roles attested in the majority of JAM's 27 CS utterances addressed solely to his sister. The first example shown below is the fourth of four cases involving reformulation (see earlier discussion for the other three). In this case JAM switches to English, repeating 'okay' twice.

(151)
*JAM: <<(es)tá bom (es)tá bom>[@pt]>[//], <okay okay>[@en]. [+ pe]
%add: MEG
Okay, okay, okay, okay. F103: L151

Although this switch to English might not seem particularly significant, it does appear to signify the beginning of a noticeable increase in JAM's use of English with MEG, as illustrated in the second CS example below, recorded five weeks after their arrival:

(152)
*JAM: <the ball>[@en] <<parece (es)tá>[@pt]>[?] quite[@en] limpo[@pt] . [+ epep]
%add: MEG
The ball looks like it's quite clean. F105: L85

The siblings are washing some golf balls in a sink and JAM makes a comment about the one he is cleaning. With three words in English (a noun phrase and an intensifier) and three words in Portuguese (a verbal phrase and an adjective), it is not possible to class either language as the ML/EL. It is plausible to suggest, then, that this CS utterance might be evidence of composite code-switching, whereby both languages contribute more equally to the CS utterance. After more than a month of immersion in an English-speaking environment, one would expect English to be taking on a greater role in JAM's interactions with other bilinguals and this particular utterance may be capturing the point at which Portuguese seems to be on an equal footing with its rival.

However, in the final example of JAM's code-switching with his sister, it appears that English has finally gained the upper hand. Occurring two months after their arrival, JAM is at the dinner table with MEG (and MOT and PAI) and is talking to

her about his icecream.

(153)
*JAM: o(lha)[@pt] <got one of the frozen bits>[@en] . [+ pe]
%add: MEG
Look, got one of the frozen bits. F111: L1618

With the only Portuguese contribution to this CS utterance being a token 'olha', there appears to be little doubt that here English has relinquished Portuguese of its role as the Matrix Language. In order to lend further support to this interpretation (based on a single CS utterance), I referred back to the relevant transcription and discovered that most of the exchanges between the siblings were actually in monolingual English. A word frequency analysis of all of the English and Portuguese tokens the siblings addressed to each other in this particular file[180] confirmed the dominance of English: for JAM the total token count for English was 285 as opposed to only 8 tokens for Portuguese; for MEG the totals were 531 and 28 respectively. These totals are in marked contrast to those found for another family meal time interaction (involving all four bilinguals) recorded a month before leaving Brazil[181]: in File 097 the Portuguese tokens occurring in exchanges between the siblings numbered 46 for JAM and 150 for MEG while the token count for English was zero for JAM and only 4 for MEG. Although these numbers are lower than those found in File 111, they still support the notion of a change in language dominance in the interactions between the siblings over a period of less than four months: before moving to England, Portuguese was evidently the more normal form of communication between JAM and MEG; a little over two months after the move, English had come to play a more dominant role. And despite the relatively low number of JAM's CS utterances available for analysis (27), this change in dominance is reflected in the differences in how English and Portuguese contribute to these bilingual utterances, as discussed above.

7.2.2 MEG's code-switching with her brother

[180] kwal @ +t%add +t*JAM +s"MEG" +u +d | freq +o -s"@nonwords.cut" -s"<@pt>" +r5
kwal @ +t%add +t*JAM +s"MEG" +u +d | freq +o -s"@nonwords.cut" -s"<@en>" +r5
kwal @ +t%add +t*MEG +s"JAM" +u +d | freq +o -s"@nonwords.cut" -s"<@pt>" +r5
kwal @ +t%add +t*MEG +s"JAM" +u +d | freq +o -s"@nonwords.cut" -s"<@en>" +r5

[181] The frequency analyses were the same as those in the footnote above, the only difference being in the selection of the file.

When it comes to examining MEG's code-switching with JAM, we are faced with an even more reduced number of CS utterances for analysis: out of a potential 34, only 14 of these are addressed soley to JAM. Ten involve the insertion of single typical Embedded Language items in English such as nouns (beyblade, plug, boot and ham), adjectives (invisible x 2) and discourse markers (no x 2 and look x 2). The ML/EL asymmetry is clearly evident in these utterances and no further discussion of them is necessary. The remaining four CS utterances, however, merit commentry and will be discussed in chronological order, starting with the one shown below:

(154)
*MEG: +" <I'm>[@en] James@pn, <eu estou pronta>[@pt] . [+ ep]
%add: JAM
I'm James, I'm ready. F011: L110

The symbol at the beginning of this utterance (+") tells us that MEG is quoting speech and although in section 6.5.2 we saw that MEG makes use of this function of code-switching, it does not explain why she should code-switch *within* the actual quote. Her use of 'I'm' and then the Portuguese equivalent 'eu estou' does not appear to be a case of retracing or reformulation (as indicated by the absence of the [//] and [///] symbols). On examination of this utterance within the transcript we learn that MEG and JAM are preparing two dolls (Barbie and Barney) for an imaginary party and that MEG is quoting JAM's doll as being ready to go. The utterance shown above is immediately followed by a monolingual Portuguese quote, 'MEG: +" <eu também>[@pt]. ('Me too') which she addresses to JAM who then responds with his own quote in Portuguese. There then follows an extensive conversation between MEG and her MOT instigated by MEG's question as to whether she should use English or Portuguese when quoting the dolls. Despite MEG's insistence on MOT making this decision, rather than suggest the language, MOT encourages MEG to think about the nationality of the doll (i.e. where it was bought) and her physical attributes (i.e. her hair and eye colour). The resulting discussion extends for over 40 utterances and concludes with the agreement that both dolls should speak English due to where they came from. However, neither doll has the opportunity to 'speak' as both JAM and MEG become preoccupied with other issues such as the age of their dolls, the party music and their desire to eat 'rapadura' (sugarcane sweets)!

By analysing the linguistic context in this way, the interpretation of the above CS utterance has been greatly enhanced and has revealed more about MEG's

language awareness. The fact that she is clearly not satisfied with her doll's use of both English and Portuguese, that is, code-switching, appears to be a reflection of her own professed desire to avoid code-switching, which she has directly and indirectly manifested in some interactions.

It is important to highlight that the remaining three example CS utterances MEG addresses to JAM all occur in England, the first two while on holiday and the third one almost a month after the siblings' move to England. This immediately implies a potential influence of the linguistic environment on the nature of MEG's code-switching, evident in the first case below (recorded 3 weeks after their arrival):

(155)
*MEG: he's[@en] <cinza>[@pt] . [+ ep]
%add: JAM
He's grey. F058: L545

Despite consisting of only two words, the grammatical nature of 'he's' places English as the ML in this CS utterance, which is contrary to the normal pattern. However, when we look at the transcript we find that both children are involved in colouring activities where their mother and English Grandmother (GRA) are actively present. It is plausible to suggest that the English environment has triggered 'he's'. When, in her next utterance, addressed to JAM and MOT, MEG then supplies the English 'grey', we are able to see her desire to accommodate to the language environment of those adults present .

In the next example MEG is reading an English story about steamrollers to MOT in the bedroom and JAM is the only other person present. JAM is unable to see the book and insists on being shown the picture of the steamroller. MEG retorts that the page she is currently on does not contain a steamroller:

(156)
*MEG: <the steamroller>[@en] <não está aqui>[@pt] . [+ ep]
%add: JAM
The steamroller is not here. F068: L866

It is unlikely that MEG knows the Portuguese word for steamroller and so its use here is not surprising. What is perhaps surprising is that MEG manages to suppress a total switch to English. Two months into their holiday and engaged in reading an English story aloud, it would have been easy to have continued with 'is not here'. However,

she sticks to her more normal pattern of code-switching when addressing JAM, with Portuguese as the ML.

The final example shown below is taken from another game of 'Guess who?' and shows MEG using English to initiate her yes/no question to JAM.

(157)
*MEG: <has he got>[@en] <cabelo amarelo[*]>[@pt] ? [+ ep]
%add: JAM
Has he got yellow hair? F103: L636

Although it would be reasonable to simply assume that a month's immersion in English has triggered MEG's use of 'has he got', an examination of the transcript of this particular game of 'Guess who?' provides evidence which implies that this code-switch has a more marked use. Occuring towards the end of the recording, MEG (and JAM) had already asked this question at least 5 times in previous rounds of guessing – in monolingual Portuguese. In fact all of the exchanges between the siblings throughout the game had been in monolingual Portuguese. In the lead up to the CS utterance shown above, MEG first asks the question in monolingual Portuguese (line 620)[182]. JAM is not ready to answer (he is busy putting down some faces on his board) and MOT tells MEG to wait. Impatient, MEG then repeats the question (in Portuguese) to JAM (line 628) and implores her mother to 'just say' ('yes' or 'no'). Told to wait again, MEG then appears to employ different tactics and code-switches (see the CS example above). Rather than being the result of the influence of the linguistic environment, the evidence suggests that this switch carries purpose and is an attempt to finally get a response to her yes/no question from JAM (and MOT), which she does so immediately.

Again we see how important it is to be able to examine the individual CS utterances (resulting from the KWAL analyses) within their linguistic context. Although, at times, this may mean reading through the whole transcript in order to understand why a particular code-switch occurs (as was the case above with this last CS utterance), the format of a CHAT transcription makes this an easy task. Unfortunately, this is not always the case with electronic corpora, many of which are designed to be machine-read only. Furthermore, the availability of my corpus means

[182] Although in both lines 620 and 636 'amarelo' ('yellow') is marked as an error, it is important to point out that in line 51 of the transcript, MEG had already explained to her mother in an aside that 'amarelo escuro e amarelo é loiro["]' ('dark yellow and yellow are 'blond"'). Aware of the correct Portuguese term from the outset, MEG nevertheless uses the incorrect 'amarelo', perhaps as a way of faciliating her brother's comprehension and thereby speeding up the game.

that any of the interpretations given above (and indeed throughout the whole of this study) can easily be critically re-examined by other researchers as they have immediate access to the primary data.

Returning to my analysis of MEG's 14 CS utterances addressed to JAM, with the exception of the four examples discussed above, her code-switching is clearly of the classic type with Portuguese acting as the Matrix Language and English taking on the more limited role of the Embedded Language. Similar roles were also found in JAM's data although English does appear relatively more frequently, giving rise to 27 CS utterances in total. The analysis and discussion of those utterances which seem to differ from this ML/EL asymmetrical pattern have revealed the linguistic environment to be a major factor in determining the structural outcome of the siblings' bilingual utterances addressed to each other. There is even evidence of a changeover in ML/EL in one of JAM's CS utterances (see above) occurring two months after the siblings' permanent move to England. It was in this file (File 111) that frequency analyses suggested that English had taken over as the main form of communication between the siblings.

The nature of the remaining files in the corpus (Files 112-119) means that there is no more attested CS data between the siblings to analyse. While Files 112 and 113 contain recordings of the siblings reading English stories to their mother, Files 114 to 119 all feature telephone interactions with relatives in Brazil. In fact, a frequency analysis of this group of 8 files[183] revealed that JAM addressed a total of only 12 words to MEG (5 English and 7 Portuguese) while MEG's total to JAM was a mere 4 words of Portuguese! Had there been further recordings of meal time interactions, it might have been possible to have confirmed that English was indeed replacing Portuguese in the siblings' monolingual exchanges and that the former had taken over as the Matrix Language in any of their bilingual utterances[184].

The discussion in this section shows how important it is to be able to examine the corpus data from different perspectives. The indepth utterance-level analysis of the code-switching occurring between the siblings has revealed a ML/EL asymmetry that was not evident in the quantitative analyses of the same data presented and

[183] See footnote 180 for the command lines used.

[184] As mentioned in section 3.1, further recordings were carried out from February 2007 (just over 2 years later) and these include meal time interactions and a game of 'Guess who?'. Although clearly such material would be ideal for comparative purposes, at the time of writing this dissertation these recordings remain untranscribed.

discussed in Chapters 4 and 5. These frequency analyses had suggested that English and Portuguese were contributing more equally in terms of the structuring of the siblings' CS utterances when addressing each other. I would like to posit that the reasons for these different findings are methodological, as explained below.

First, as pointed out previously, when we analyse the data as a whole, any differences in code-switching patterns across the data (resulting from the influence of contextual and longitudinal factors) are necessarily 'averaged out'. The effect this would have on the quantitative results was discussed in section 4.1.4.

The second methodological issue is related to the use of the addressee tier when specifying the input for any of the CLAN analyses. From the utterance-level analyses it has become clear that when one specifies the addressee, CLAN will select any utterances where the target addressee code is found, including those which are multi-addressed. When examining the KWAL output qualitatively (see current and previous chapters), it is possible to take into account the influence of these other addressees and, if so desired, exclude certain utterances from the analysis (as in 7.1 and 7.2). However, when carrying out the quantitative analyses and the word/code-level analyses (Chapters 4 and 5) such exclusion did not occur. This means that the input will have included data which is multi-addressed. Clearly, in the case of the siblings' interactions with each other, the inclusion of such data has led to a certain skewing of results and explains why the quantitative results do not reflect what has been uncovered by the subsequent qualitative analyses. Indeed, if we consider that the majority of the excluded multi-addressed CS utterances (21 out of 48 for JAM and 20 out of 34 for MEG) had their mother as the other addressee, then this would account for the more balanced use of English and Portuguese apparent in the quantitative results, the majority of the English tokens actually due to MOT's presence as an additional interlocutor in the interactions. This methodological issue does not seem to have presented such a problem in the analysis of other speaker/interlocutor combinations (such as JAM/MOT, MEG/JAM, JAM/PAI and MEG/PAI): this is shown by the fact that the utterance-level analyses support what was suggested by the quantitative results for these speakers.

From the discussion above it has become evident that the automatic inclusion of multi-addressed utterances in the input for specific speaker/interlocutor combinations is unsatisfactory. In the case of my study where an additional addressee can affect a speaker's code-switching patterns, this methodological

problem needs resolving if quantitative results are to be reliably interpreted. As this issue is crucial to my methodology, I will return to this discussion in Chapter 8 where I present a solution to the problem. For now, however, I will continue with my utterance-level analyses, this time turning to the code-switching occurring between the parents where an examination of even a very reduced number of utterances proved to be very enlightening.

7.3 Code-switching between the parents

Despite the very little amount of attested code-switching which took place between the siblings' parents, the analysis of the CS utterances resulting from their KWAL analyses[185] still proved to be fruitful, especially the analysis of MOT's utterances addressed to PAI.

With regards to PAI's code-switching with MOT, there were only 6 CS utterances in the output, one of which I excluded from the analysis as it was also addressed to his children. The remaining 5 utterances, presented in chronological order below, reflect a classic asymmetry in the roles of Portuguese and English, the former as the Matrix Language and the latter as the Embedded Language contributing with single items (boarding, single, packed lunch, so and witness):

(158)
*PAI: <não é>[@pt] boarding[@en] <não, ele falou o portão c@l>[@pt] . [+ pep]
%add: MOT
It's not boarding, no, he said gate C. F52: L423

(159)
*PAI: <esse duvet é duplo o é>[@pt] single[@en] ? [+ pe]
%add: MOT
Is this duvet double or is it single? F111: L 331

(160)
*PAI: +< <vai ter uma coisa para o meu>[@pt] <packed lunch>[@en] <amanhã>[@pt] ? [+ pep]
%add: MOT
Is there going to be something for my packed lunch tomorrow? F111: L641

(161)
*PAI: <so>[@en] <tu acha que eu devo voltar pela[//] pelo meio da rua principal>[@pt] ? [+ ep]
%add: MOT

[185] kwal @ +t%add +t*MOT +s"PAI" +u +s"[+ *]" and kwal @ +t%add +t*PAI +s"MOT" +u +s"[+ *]"

So, you think that I should come back by the, by the middle of the high street?
F111: L1037

(162)
*PAI: <tem um>[@pt] witness[@en] <se não ninguém acredita em mim>[@pt] . [+ pep]
%add: MOT
There's a witness, if not nobody believes me. F111: L1480

It is relevant to note that all but one of the utterances (the first one) occur in the same file (111), a meal time interaction recorded after the family's move to England. With only one attested CS *before* the move it does seem plausible that PAI's language use is being affected by his new linguistic environment. This increase in use of English EL items is not surprising and was attested in the CS utterances of both JAM and MEG when addressing their father over the telephone while in England (see section 5.1.2). Based on these findings one might expect that *MOT*'s already restricted use of Portuguese in her CS utterances would be further diminished as English took over as the dominant language of the family's environment. However, as will be seen below, her CS utterances revealed something rather unexpected.

Despite there being only 6 CS utterances which MOT addresses to PAI, when one compares the participation of English and Portuguese across the six utterances, it is possible to note something potentially significant. In the first three CS utterances (shown below), it is clear that MOT is using English as the ML, with Portuguese contributing single EL items (centavos, olha and padaria). This pattern is unsurprising given that such classic asymmetrical use was also found in MOT's interactions with her children.

(163)
*MOT: <seventy one>[@en] <centavos>[@pt] <for two like that>[@en] . [+ epe]
%add: PAI
Seventy one cents for two like that. F079: L502

(164)
*MOT: <o(lha)>[@pt] <James@pn has a practice tomorrow>[@en] . [+ pe]
%add: PAI
Look, James has a practice tomorrow. F089: L235

(165)
*MOT: <I'm going to the>[@en] padaria[@pt] . [+ ep]
%add: PAI
I'm going to the baker's. F098: L168

However, when we examine the remaining three CS utterances, we no longer find this pattern. Indeed, the language roles appear to have switched, with English contributing only single EL items (bill and Alex[186]) and a quoted phrase (from a telephone message):

(166)
*MOT: <mas eu não sei quanto tempo[///] porque é[/] é cada três meses que minha Mãe@m recebe um>[@pt] bill[@en] . [+ pe]
%add: PAI
But I don't know how long, because it's it's every three months that my mother gets her bill. F104: L769

(167)
*MOT: <Alex@pn>[@en] <a ligação caiu>[@pt] . [+ ep]
%add: PAI
Alex, our call got cut off. F116: L581

(168)
*MOT: <xxx eu acho que entrou no>[@pt] <<welcome to Telestunt>["]>[@en] . [+ pe]
%add: PAI
I think it went onto 'Welcome to Telestunt'. F116: L586

These typical contributions from the Embedded Language are found within a Portuguese frame, the latter now clearly in the role of the Matrix Language. Taking into account all the evidence thus far discussed in this study regarding MOT's code-switching practice, this abrupt reversal of language roles is rather unexpected. This is especially so when we learn that the latter three utterances occur in an English environment, after MOT has moved to England. One might assume that such immersion would reinforce MOT's dominant use of English in her code-switching patterns and perhaps reduce the already limited role that Portuguese has to play. However, this is clearly not the case - the opposite appears to have occurred!

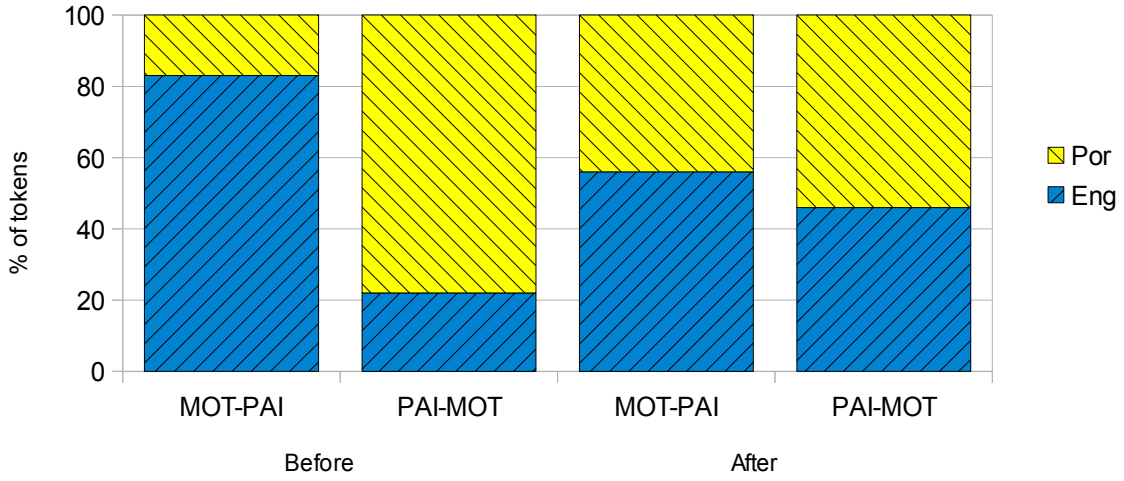
In order to further investigate this finding, I decided to examine MOT's overall use of English and Portuguese (not just in CS utterances) with her husband before and after moving to England. For comparative purposes I also looked at PAI's language use with MOT. Separate frequency analyses for each language were performed on two sets of files, set 1 containing recordings carried out before the move to England (Files 001 to 099) and set 2 consisting of those carried out after the move (Files 100 to 119). This resulted in 8 frequency analyses (2 speakers x 2

[186] MOT's use of 'Alex' has been coded as English due to the fact that she also uses Alexandre (his Brazilian name) on other occasions.

languages x 2 sets of files)[187]. As the number of files in each set are not equal (99 in set 1 and only 20 in set 2), rather than present a graph showing the raw frequency data for each set, I have converted the results for each language into percentages – this allows for more insightful comparisons across the different results.

If we look at the results in Fig. 24 for the period of time *before* the move to England, it is possible to see the dominance of the respective mother tongue in all exchanges between MOT and PAI: 83% (1002 tokens) of MOT's tokens are English while 78% (764 tokens) of PAI's tokens are Portuguese.

Figure 24. Proportions of English and Portuguese tokens exchanged between MOT and PAI before and after moving to England



With only 17% (207 tokens) of MOT's total number of tokens consisting of Portuguese, this would indicate that MOT is not allowing her Portuguese language environment to impact too much on her use of English with PAI. This interpretation is supported by other findings which showed MOT's very restricted use of Portuguese with her children for this period: only 1% of the tokens she addresses to both JAM and MEG are Portuguese[188]. If we compare the proportion of MOT's use of her second language when addressing PAI (17%) to PAI's use of his second language (English) with MOT (22%), we have evidence to suggest that MOT is restricting the

[187] kwal @ +t%add +t*MOT +s"PAI" +u +d | freq +o -s"@nonwords.cut" -s"<@pt>" +r5
kwal @ +t%add +t*MOT +s"PAI" +u +d | freq +o -s"@nonwords.cut" -s"<@en>" +r5
kwal @ +t%add +t*PAI +s"MOT" +u +d | freq +o -s"@nonwords.cut" -s"<@pt>" +r5
kwal @ +t%add +t*PAI +s"MOT" +u +d | freq +o -s"@nonwords.cut" -s"<@en>" +r5

[188] The first two analyses in the footnote above were repeated with JAM and MEG as addressees (+s"JAM" and +s"MEG") after first selecting files 001 to 099.

use of her second language more than her husband. This 5% difference is even more significant if we take into account the potential influence of the linguistic environment which would favour the use of Portuguese and not English: despite being immersed in Portuguese, 22% (218 tokens) of PAI's tokens are still English. All of this points to a conscious decision on MOT's part to use her mother tongue as much as possible with her husband and her children.

For the period of time after the family's move to England, we can observe a change in language use between MOT and PAI. For PAI this change involves an increase in the proportion of English tokens he addresses to MOT when compared with the previous period: from 22% to 45% (193). As only 5 of these English tokens occur in CS utterances (as seen in the ones discussed above), the remainder of these tokens must occur in monolingual English utterances. There can be little doubt that this increase in PAI's use of English is the result of the influence of the linguistic context.

Based on this evidence one would therefore have expected that the proportion of English tokens MOT addresses to PAI would also have increased. This is especially so if we consider that in the period before moving to England MOT appears to make a concerted effort to use as much English as possible with PAI (and JAM and MEG). However, it is clear from the chart that there is actually a *de*crease in the amount of English she uses with her husband: the percentage of English tokens falls from 83% to 56% (314). This means that she is thus using more Portuguese with her husband in England than she did while living in Brazil: now 44% (243) of her tokens are Portuguese, 25 of these occurring in 3 CS utterances (as shown above). In order to see whether she is also using relatively more Portuguese with her children, 4 more frequency analyses were carried out on her exchanges with both siblings for the second set of files[189]. Although the results revealed only a slight increase in the amount of Portuguese addressed to JAM and MEG (from 1% to 4% for both children), the fact that MOT is actually using more Portuguese is significant given the linguistic context.

Such findings only emerged following further investigation into an unexpected pattern of CS use found in 3 of MOT's 6 CS utterances addressed to PAI. By performing subsequent frequency analyses and cross-referencing the results with contextual information it has been possible to gather evidence to support the notion

[189] See footnote 187 but note that this time Files 100 to 119 were pre-selected.

that MOT's language use with her husband and children is influenced by the linguistic environment but a different way. While in Brazil she appears to have made a conscious decision to use English as much as possible, in England (after the move) it appears she has decided to make more use of Portuguese, especially with her husband. This change in language strategy might reflect MOT's desire to ensure that her children continue to be exposed to Portuguese while immersed in an English context. With no source of Portuguese outside the home environment, there appears to be a realisation that it is up to them (PAI and MOT), to provide such exposure. One could interpret from the analyses presented here that MOT appears to be more committed than PAI to implementing this new language strategy. Whether this is actually true cannot be ascertained but the analysis of PAI's own code-switching with his children, presented below, may reveal more.

7.4 PAI's code-switching with his children

The final set of CS utterances to be analysed are those addressed by PAI to his children. From previous quantitative analyses we already know that the data for PAI is very reduced mainly due to his absence in most of the interactions that were recorded. Although he features heavily as an interlocutor in the telephone interactions, as his turns were not recorded we do not have access to these particular utterances. However, a brief look at the KWAL output of his code-switching with JAM and MEG[190] does highlight a particular function of his restricted bilingual language use, as will be seen below.

With the removal of those CS utterances addressed to more than one speaker (3), we are left with only 3 that PAI addresses solely to JAM and 6 (out of 9) that are addressed to MEG. From the metalinguistic coding found in each of the 3 CS utterances addressed to JAM (shown below), it is clear to see that PAI's code-switching in each case is the result of the quoting of entire phrases.

(169)
*PAI: <a[@en] pouco[@pt] <of mine>[@en]>["][= laughing] . [+ epe]
%add: JAM
'A little bit of mine'. F015: L401

(170)

[190] Resulting from the following two analyses: kwal @ +t*PAI +t%add +s"JAM" +u +s"[+ *]" and kwal @ +t*PAI +t%add +s"MEG" +u +s"[+ *]"

*PAI: <just[@en] <um pouco>[@pt]>["] ? [+ ep]
%add: JAM
'Just a little bit'. F015: L429

(171)
*PAI: <<an evening with the Keiths@pn>["]>[@en] rapaz[@pt] ! [+ ep]
%add: JAM
'An evening with the Keiths', my boy! F111: L1060

An examination of File 015 (a meal time interaction) revealed that PAI is actually quoting JAM's use of a code-switch with his mother which involved him offering MOT 'um pouco' ('a little bit') of his own juice. Initially PAI finds JAM's use of 'um pouco' within an otherwise English utterance amusing (see first example above). However, when JAM goes on to use it twice more with his mother, PAI then quotes him again, his rising intonation indicating his questioning of this usage (second example above). The fact that PAI then immediately offers JAM a translation into English (PAI: <<just a little bit>["]>[@en]. – L431) could be interpreted as revealing a preocupation with JAM's 'unnecessary' use of Portuguese. However, there is no response from JAM to his father's intervention: the former seems more preoccupied with his food and the fact that he was eating more than his sister.

In the third utterance shown above (171) PAI is responding to JAM's question to his mother about where his father is going (PAI is getting ready to go out). He quotes the title of the photographic society talk he is going to before ending with a rather emphatic 'rapaz!', reflecting his disbelief that JAM had not paid any attention to some of the meal time interaction which had revolved around the discussion of PAI's evening out.

With only three attested CS utterances to analyse, it is impossible to draw any conclusions about PAI's use of code-switching with his son. However, an examination of the particular interactions from which they were taken have provided us with an insight into what lies behind PAI's use of the quoting function discussed above – his metalinguistic awareness of JAM's code-switching habits. This awareness, and desire to raise his children's awareness, is also evident in five of the six CS utterances PAI addresses to MEG which are discussed below.

The first two CS utterances shown occur during the recounting of an incident in which JAM broke a panel of glass and MEG got some glass on her toe. In the first instance, while addressing her MOT, MEG code-switches and uses the Portuguese

word for glass (vidro) twice before PAI steps in and asks her to choose between the Portuguese or the English word:

(172)
*PAI: <vidro["] ou>[@pt] <glass["]>[@en] ? [+ pe]
%add: MEG
'Glass' or 'glass'? F015: L372

He immediately then asks a second question, <are you talkin(g) English or Portuguese>[@en]? (L374) in order to clarify what he implied by the first question. MEG laughs and replies with the English alternative. His intervention has had the desired effect of reminding MEG that she 'should' be using English with her mother. It is interesting to note that PAI himself asked the second question in monolingual English! Whether this was a conscious decision or not is difficult to say but it may have had a priming effect on MEG's choice between the two alternatives as she promptly chooses 'glass'. After MEG's response, PAI then appears to quote MEG's original code-switch although the utterance is incomplete (the symbol xxx meaning that part of it was unintelligible[191]):

(173)
*PAI: <<xxx some[/] some>[@en] vidro[@pt]>["] +... [+ ep]
%add: MEG
'*Some, some glass*'. F015: L379

In the third of the CS utterances (see below), PAI quotes MEG as saying 'terra' ('ground'). This time, instead of representing a 'criticism' of her use of a Portuguese word while playing an animal guessing game with her mother, he appears to be simply pointing out that MEG had not specified if the animal she is describing lives *in* or *under* the ground[192].

(174)
*PAI: <you just said>[@en] <terra["]>[@pt] . [+ ep]
%add: MEG
You just said 'ground'. F028: L208

Again, in this instance, PAI is unexpectedly using English as the Matrix Language with MEG. As in the interaction discussed previously, it is plausible to suggest that

[191] See Appendix B for more details of these transcription conventions
[192] In actual fact MEG had already used the word 'underground' (L141) before her use of 'terra' (167) but MOT's question, MOT:<where does it live>[@en]? (L165) clearly shows that the latter had not heard MEG's earlier use of the English term.

PAI is accommodating his language use to the pattern of mostly monolingual English exchanges occurring between mother and daughter as they play the game.

The following fourth CS utterance occurs as the family are eating pizza. Most of the entire conversation in this recording revolves around pizza and at one point MEG declares that if she worked in a pizza shop she would be the 'pizza eater' (L381). PAI picks up on this later and uses it in his question to her:

(175)
*PAI: <Meggie@pn teu próximo dia de natal queria ser um>[@pt] <<pizza eater>[@en]>["] ? [+ pe]
%add: MEG
Meggie on your next Christmas day would you like to be a 'pizza eater'?
F050: L587

Whereas in the above CS utterance PAI is simply quoting his daughter's coined usage of 'pizza eater', in the fifth utterance below PAI's quoting is more metalinguistic in nature as he explains to MEG the meaning of the Spanish word 'burro', used in a quote from Fawlty Towers:

(176)
*PAI: <porque no espanhol>[@pt] <burro["]>[@sp] <não é burro["], é>[@pt] <donkey["]>[@en] . [+ pspe]
%add: MEG
Because in Spanish 'donkey' is not 'stupid' it's 'donkey'. F079: L742

As explained in the section on the siblings' use of Spanish words, much of the humour in Fawlty Towers revolves around linguistic misunderstandings between Manuel, the waiter, and Basil, the hotel owner. MEG is attempting to understand the humour behind Manuel's quote and asks her mother <why did he say that Manuel@pn>[@en] <<but a>[@en] burro[@sp] <is a eeyore>[@en]>["]. [+ esse] (L737). MEG's understanding is complicated by the fact that 'burro' in Brazilian Portuguese (at least in the North-East region) means 'stupid', there being a different word for 'donkey' ('jumento'). It is PAI who offers an explanation (see above), quoting first a Spanish word ('burro')[193], a Portuguese word ('burro') and then an English word ('donkey'). Despite this multiple language quoting, PAI's CS utterance is clearly framed by Portuguese, which is acting as the Matrix Language.

[193] The Spanish pronunciation of 'burro' is ['burɔ] as opposed to ['buhu] in Portuguese.

PAI's use of Portuguese as the ML can also be seen in the final CS he addresses to MEG. In England, talking about the programmes on television that evening, PAI tells MEG there are two nature programmes.

(177)
*PAI: <tem um o dois programas de>[@pt] nature[@en] .[+ pe]
%add: MEG
There's one or two nature programmes. F111: L448

It is unsurprising that PAI inserts the English word nature, especially if one considers that the TV guide consulted is in English. When he then goes on to give more details about the programmes in entirely monolingual English utterances, it is possible to see the influence of the linguistic environment on his language use.

This last CS utterance stands out as being the only one where PAI does not appear to be overtly code-switching for the purposes of quoting: in all of the other CS utterances addressed to JAM and MEG (8 in total), code-switching has served the functions of quoting and metalinguistic referencing. It seems likely that, now in England, both the frequency and functions of code-switching employed by PAI would increase as English comes to play a greater role in his every day language use.

Although there is no more attested CS data to confirm this supposition, frequency analyses of his use of both English and Portuguese with JAM and MEG before and after the family's move to England[194] do provide evidence for a change in the language balance. In the results for JAM as addressee there is an increase in PAI's use of English from 17% to 39% and for MEG as addressee the increase is from 23% to 27%[195]. While for both children there is evidence that PAI is using more English after the move, the differences in percentages reveal something notable: while PAI uses only 4% more English tokens with MEG after the move, with JAM there is a 22% increase in English tokens. If we consider that the family share the same linguistic environment, it is not possible to account for these differences solely in terms of the influence of the English environment on PAI's use of English. It may be that PAI is accommodating his language use according to changes in how his

[194] kwal @ +t%add +t*PAI +s"JAM" +u +d | freq +o -s"@nonwords.cut" -s"<@pt>" +r5
kwal @ +t%add +t*PAI +s"JAM" +u +d | freq +o -s"@nonwords.cut" -s"<@en>" +r5
kwal @ +t%add +t*PAI +s"MEG" +u +d | freq +o -s"@nonwords.cut" -s"<@pt>" +r5
kwal @ +t%add +t*PAI +s"MEG" +u +d | freq +o -s"@nonwords.cut" -s"<@en>" +r5

[195] The raw frequency data for both sets of files were as follows: in Files 001-099 the numbers of tokens for English and Portuguese were 120/607 with JAM as addressee and 171/584 with MEG as addressee; in Files 100-119 the total were 118/188 and 105/286 respectively.

children are addressing him. Indeed further frequency analyses[196] do suggest that this might be the case: after the move JAM's use of English with this father increases from 13% to 16% while MEG's use of English with PAI actually *de*creases, from 8% to 7%. This decrease might have had the effect of restraining PAI's use of English with MEG (which increased by only 4%) while JAM's 3% increase in English may have encouraged PAI to use more English in return. Although this data does appear to suggest a certain language repriocity between PAI and his children, it cannot totally explain the differences found: PAI's 22% increase in English when addressing his son far surpasses JAM's 3% increase when addressing his father. Therefore, accommodation is likely to be only one factor affecting PAI's increasing use of English with his children.

The discussion in this section has shown how fruitful the investigation of only a few CS utterances can be. From the original KWAL output of 9 CS utterances addressed to JAM and MEG, I was able to search for explanations for PAI's bilingual usage by consulting the files and by performing further frequency analyses. My findings revealed that PAI's code-switching with JAM and MEG served very specific purposes, that of highlighting his children's own code-switching practices. With such overt attention to the siblings' bilingual language use also evident in MOT's data (see 7.1), it would not be surprising to find a more heightened language awareness in both JAM and MEG when compared to monolingual peers. It is clear from the analyses of MOT's, and now PAI's data, that they both have key roles to play in directly and indirectly affecting their children's code-switching practices and more general language use. It would be of great interest to carry out a more systematic examination of the parents' pragmatic reactions to their children's code-switching and the subsequent influence that different strategies have on the siblings' bilingual practices. I could use Lanza's categorization of parental discourse strategies framework (1997, 2007) to analyse the strategies used by MOT and PAI and compare my findings to those reported on in the longitudinal study carried out by Juan-Garau and Pérez-Vidal on their bilingual (Catalan/English) son Andreu (2001). Although such an indepth analysis is beyond the remit of the current study, it remains earmarked for future investigation.

[196] The speaker and addressee codes in the four command lines in footnote 194 were simply switched to carry out these analyses. For example, +t*PAI +s"JAM" became +t*JAM +s"PAI" etc.

The analyses in this chapter focussed on the code-switched data of MOT and PAI and on the CS utterances exchanged between the siblings. Despite there being reduced numbers of utterances to examine, I hope to have shown in my discussions of the examples how enriching such a qualitative analysis can be. Such discussions have also served to highlight the overarching importance of taking into account information regarding context and addressees when interpretating code-switched data. The issue of multi-addressed utterances was first brought to the fore in section 7.2 when I examined the CS utterances exchanged between the siblings. Although the exclusion of utterances addressed to more than one person did not present a methodological problem in the current qualitative analysis (I was able to exclude them mannually), it was clear that to effect such exclusion in my quantitative analyses a different methodological approach would be needed. This approach is discussed in detail in section 8.5 of the following, and final, chapter in which I consider the implications of my investigation of code-switching in the LOBILL Corpus.

8. Conclusions and implications for research in code-switching

As discussed in 2.3, through the current study I proposed to offer a three-fold contribution to the research field of code-switching: original corpus data, original methodology and original results.

In terms of the corpus itself, its contribution to the CHILDES data base means that original naturalistic data for the language pair Portuguese/English is made available to the wider academic community for further linguistic enquiry and for cross-linguistic comparative research. As has been demonstrated in this study, the heterogenous and longitudinal nature of the LOBILL Corpus makes it a particularly rich source of linguistic enquiry and its specific language coding significantly increases its exploitability.

In terms of methodology and results, throughout this dissertation it was demonstrated how the different codes inserted in the corpus (particularly the language codes, the CS postcode and the addresse codes) made it possible to perform a myriad of analyses via the CLAN tools. Chapter 4 reported on the quantitative analyses (via the commands FREQ, VOCD and WDLEN) which provided empirical results to support my hypotheses which proposed relationships between four types of values and the contributory roles of the languages in CS utterances. The discussion of the word- and code-level analyses of the data in Chapter 5 revealed how fruitful such an approach could be in terms of investigating the relationship between certain linguistic phenonema (such as retracings and reformulations, errors, tag questions and metalinguistic usage) and code-switching. The foci of Chapters 6 and 7 were the more qualitative analyses of code-switched data in the LOBILL Corpus which allowed not only further investigation of purely linguistic aspects of the code-switching but also made it possible to uncover the sociolinguistic and pragmatic motivations underlying the informants' use of code-switching.

By providing sufficient details about the innovative methodological approach used in this study (such as the construction of the specific command lines), the replication of my analyses is made possible. Apart from ensuring transparency and thereby enhancing the validity of my own results, the ability to replicate my analyses also means that such methodology could easily be applied to other suitably coded data sets. Subsequent comparisons across different corpora in terms of results would

no doubt offer the potential for a more enhanced investigation of the relative roles typological, sociolinguistic and idiolectal factors have to play in code-switching behaviour.

Although evidence for the contributory claims made above may permeate the entire dissertation, it is useful to highlight some of the more important theoretical and methodological implications of this study. To begin with I will discuss my proposal for a schema designed to assist in the interpretation of four types of quantitative values when used to measure the differential contribution of languages in CS utterances (8.1). I will then detail the theoretical contributions my word and code-level analyses offer the investigation of code-switching (8.2) before summarising what my utterance-level analyses reveal about the code-switching behaviour of the siblings when addressing their parents (8.3). By comparing the bilingual language use of the siblings I am able to comment on the impact of both contextual and developmental factors on their code-switching behaviour. The value of also being able to examine the CS utterances of other speaker/interlocutor combinations (in Chapter 7) will become evident in the discussion in 8.4. After then discussing specific methodological issues relating to the inclusion/exclusion of addressee tiers (first raised in 7.2) and proposing a solution to the problem (8.5), I will end by considering the implications of my study for the future of code-switching research (8.6).

8.1 Using quantitative measures to investigate the relative roles of languages participating in CS utterances

If we recall, according to the Asymmetry Principle (Myers-Scotton, 2009:209) the abstract morphosyntactic frame of a bilingual clause largely, or entirely, comes from the Matrix Language while the Embedded Language (EL) typically contributes content morphemes, such as nouns, lexical verbs and adjectives.

In quantitative terms, I proposed it was possible to make certain assumptions regarding how the ML/EL asymmetry is realised in code-switched utterances. Firstly, it is reasonable to propose that the ML would contribute more words to a code-switched utterance than the EL. Secondly, the grammatical nature of the words contributed by the ML and their repetitive frequency would mean that there is less diversity in their contribution to code-switched utterances than the lexically-laden content morphemes being inserted by the EL. And thirdly, if one considers that in many (European) languages grammatical morphemes are typically shorter in length

(in terms of characters) than content words, one would expect higher mean word lengths for the EL when compared to the ML.

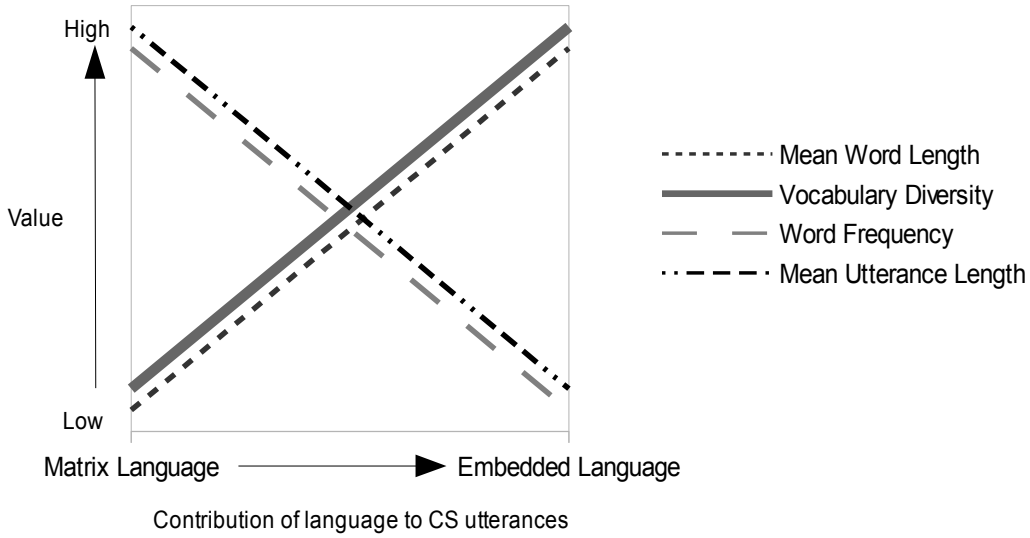
Due to the language and addressee coding in the LOBILL Corpus, it was possible to test these assumptions via the use of four CLAN commands (KWAL, FREQ, VOCD and WDLEN). As was seen in the discussion of the results of these analyses in Chapter 4, relationships were indeed found between the four different measures (word frequency, vocabulary diversity, mean word length and mean utterance length) and the participatory roles of the two languages in code-switched utterances. In their simplest from the relationships can be stated as follows:

(i) A low word frequency indicates an EL while a high word frequency indicates a ML.
(ii) A low D score indicates a ML while a high D score indicates an EL.
(iii) A low mean word length indicates a ML while a high mean word length indicates an EL.
(iv) A low mean utterance length indicates an EL while a high mean utterance length indicates a ML.

For two of the proposed relationships ((i) and (iv)), the evidence strongly suggests that the greater the relative difference in values between the languages the more asymmetrical their participatory roles appear to be. This in turn means that where the relative difference in values is less disparate we would expect to find more equal participation of the languages in code-switched utterances. As far as the other two proposed relationships are concerned ((ii) and (iii)), as will be seen in the discussion in 8.1.2 and 8.1.3, developmental aspects need to be taken into account when interpreting what the values mean in terms of the ML/EL asymmetry.

Before looking more specifically at each of the four relationships, I thought it would be useful to present a visual summary of all four combined. The following schema, therefore, represents my proposal for the interpretation of the four different types of quantitative measures at its most basic level. It highlights the notion of a continuum on which the resulting values for each language of an individual's CS utterances can be plotted and subsequently interpreted in terms of their participatory roles.

Figure 25. Schema for the interpretation of four quantitative measures when used to investigate the relative roles of languages contributing to CS utterances



If, for example, when analysing the code-switched utterances of a particular speaker, their Vocabulary Diversity (VD) score for one of the languages was found to be relatively lower than for the other participatory language, one could use the VD continuum (the solid line in the chart) to interpret what this relative difference in scores might mean in terms of the roles of both languages in CS utterances: a lower value (see the y-axis) would indicate a language acting as the Matrix Language (see x-axis) while a relatively higher value (moving up the VD continuum) would indicate a role more akin to the Embedded Language. If the values for both languages were found to be very similar, this would mean plotting both languages half way up the VD continuum, meaning, thus, that neither language could be said to be acting as the ML or the EL as both languages would then fall half way between the two extremities of the ML/EL x-axis.

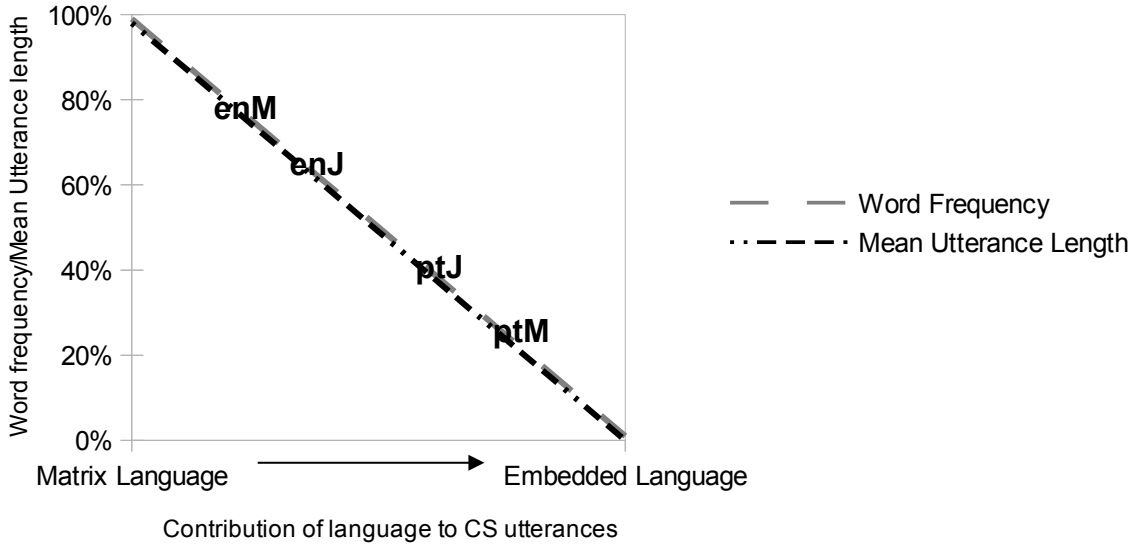
In the following sections (8.1.1, 8.1.2 and 8.1.3) I will be using specific data from my own study to offer a more detailed interpretation of each of the four continua shown in the above schema. However, I would first like to make an observation about what this visual representation suggests about the relationship between the four continua – that they seem to pair up: Word Frequency and Mean Utterance Length values follow the same predictive line while the increase/decrease in Vocabulary Diversity and Mean Word Length values follow the same pattern. Both these pairings

are logical but in different ways. First, as was pointed out at the end of 4.3.2.2, although the method used by the commands FREQ and WDLEN may be different and the output is presented differently (word frequencies as oppposed to utterance lengths), when both sets of data are converted into percentages, we arrive at the same results in terms of the proportion each language contribtes to CS tterances. With regards to the second pairing, it is logical that vocabulary diversity scores will tally with word lengths: while the ML is characterised by its high frequency of shorter grammatical morphemes (i.e low diversity and low word lengths), the EL typically contributes longer morphemes which are often singular in nature (i.e high word lengths and high diversity). This second pairing is of particular interest, as will be seen further on in the discussion when I show how my proposed schema can be used to interpret the VOCD and MWL results for MEG and JAM. In the following sub-sections I deconstruct the schema presented above in order to discuss the relationships in more detail and to be able to plot the siblings' scores more accurately on each continuum line. In each case the scores are those relating to their CS utterances addressed to MOT.

8.1.1 Interpreting Word Frequency and Mean Utterance Length scores according to the schema

In the first figure below, I have plotted the results of two of the quantitative analyses for JAM and MEG, those of Word Frequency and Mean Utterance Length.

Figure 26. Schematic representation of the relative roles of English (en) and Portuguese (pt) in JAM (J) and MEG (M)'s CS utterances addressed to MOT based on Word Frequency and Mean Utterance Length scores (converted to percentages)



By converting the original raw output of both analyses into percentages, it is not surprising to find that there is an exact match in terms of the resulting percentages of the analyses (see earlier discussion in 4.3.2.2). What is made visually clear from the plotting of the percentages for each language for both JAM and MEG is the relative roles English and Portuguese have to play in their code-switched utterances with their mother. With 76% for English and 24% for Portuguese, MEG's percentages place the contribution of each language towards the opposing ends of the ML/EL continuum, reflecting a more classic style of code-switching. JAM's percentages of 61% for English and 39% for Portuguese place each language's contribution further towards the middle of the continuum, suggesting less defined (and therefore less classic) usage of an ML and EL in his code-switched utterances.

Although I am using the schema above to interpret results of analyses performed on bilingual data, there is no reason why it would not prove useful for researchers analysing trilingual or multilingual data. For example, if one were to establish the word frequency (or mean utterance length) of a trilingual speaker's language use as being 65% English, 25% Portuguese and 10% German, these values could then be plotted on the continuum and interpreted accordingly. The same could apply to the interpretation of the other two types of values, the Vocabulary Diversity scores and the Mean Word Lengths, discussed in the following two sub-

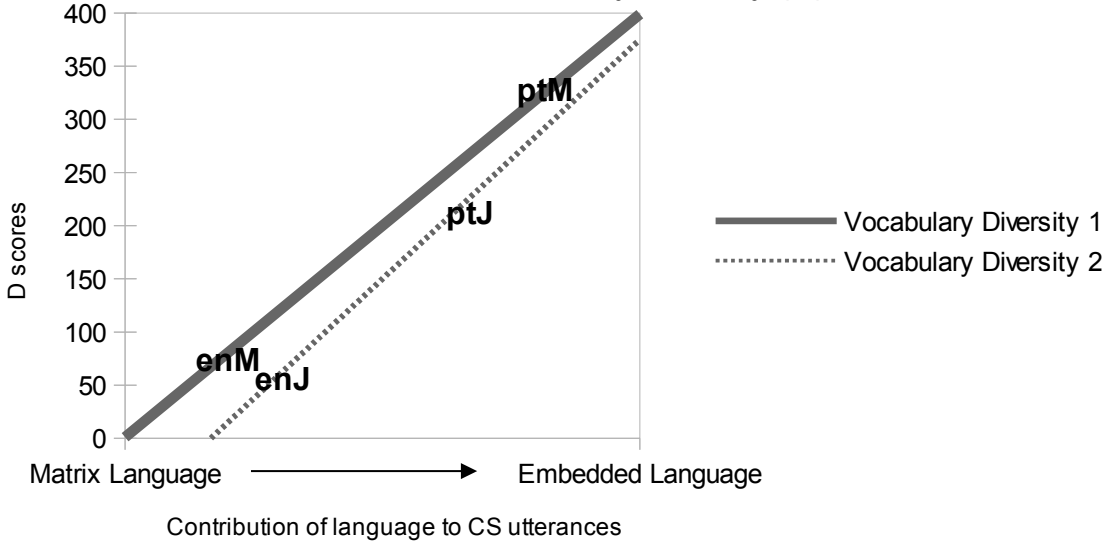
sections.

8.1.2 Interpreting Vocabulary Diversity scores according to the schema

As mentioned in the introductory part of this section, there is a need to take into account developmental aspects when using Vocabulary Diversity scores to establish the relative roles of languages participating in code-switched speech. This need first came to light in the discussion of the longitudinal analysis of JAM and MEG's D scores in section 4.2.4, where MEG's D scores for English were found to be consistently higher than JAM's. According to the original hypothesis these results would have meant that MEG's use of English as the ML was less classic than JAM's. From the triangulation of the data and the qualitative analyses of the siblings' CS utterances I knew that MEG's code-switching was more classic than JAM's and I concluded that the D scores were reflecting the fact that MEG (2½ years older than her brother) was linguistically more developed than her brother. Rather than abandon my hypothesis which proposed a promising relationship between vocabulary diversity scores and the participatory roles of the languages contributing to CS utterances, I sought to incorporate this developmental aspect in some way.

The results of experimentation can be seen in the following schema which proposes slightly different Vocabulary Diversity continuum lines for JAM and MEG.

Figure 27. Schematic representation of the relative roles of English (en) and Portuguese (pt) in JAM (J) and MEG (M)'s CS utterances addressed to MOT based on Vocabulary Diversity (D) scores



To arrive at these two predictive continuum lines I first plotted MEG's D scores (75 for English and 329 for Portuguese) on the orginal continuum (Vocabulary Diversity 1). With both scores lying towards the opposing ends of the continuum they reflect MEG's use of classic code-switching. If I were to plot JAM's D scores on this same continuum (i.e Vocabulary Diversity 1), his score for English (58) would place him to the left of MEG (implying that his use of English was more Matrix-like than his sister), while his score for Portuguese (201) would place him exactly in the middle of the continuum, which would imply that Portuguese was acting as neither the ML or the EL in his CS utterances addressed to his mother. From the qualitative analyses we know that such an interpretation would be eroneous. What was needed was a way of plotting JAM's D scores so that they would still reflect his less classic use of code-switching when compared to MEG. Experimentation led to the establishment of a second Vocabulary Diversity continuum (2) which, as can be seen above, allowed for a more accurate representation of JAM's language use when code-switching with his mother. Both of JAM's D scores are now found further towards the middle of the continuum (as in Fig. 26) , thus reflecting a less classic style of code-switching which is consistent with the quantitative and qualitative results for JAM.

The fact that the two Vocabulary Diversity continua do not run exactly parallel to each other reflects an important observation that was made when comparing the

CS D scores for JAM and MEG in each language. It was noted that the difference between the siblings' scores for English were relatively greater than those for Portuguese. With English as the Matrix Language providing most of the grammatical morphemes, it was postulated that the wider difference in D scores for English between JAM and MEG must be principally due to differences in the lexical diversity of their grammatical systems: with a more developed system, MEG would use a wider variety of grammatical morphemes than her brother. Such differences in grammatical diversity would clearly be less evident, if at all, in their D scores for Portuguese, which, acting as the Embedded Language, typically contributes content morphemes. However, one would still expect MEG's pool of content morphemes in Portuguese to be relatively larger than JAM's and that is why the two continua do not merge at the Embedded Language extreme.

Although the above schema was developed based on data for only two bilingual children, it does provide insights into how sets of D scores of other bilingual individuals could be interpreted in terms of the relative roles of the languages participating in CS utterances. It shows how it is feasible to account for both grammatical and lexical differences in vocabulary diversity which are due to age-related linguistic development. Due to the fact that the D scores for JAM and MEG represent the average of over three years of data, it would not be possible to accurately determine a particular age for each of the two continua. However, one could take the mean ages of JAM (5;1) and MEG (7;6) and use this as a baseline for comparative purposes. Based on the discussion of the present schema it would then be possible to suggest the addition of further age-related continua: for example a Vocabulary Diversity continuum for a three-year-old bilingual would lie below continuum 2 and one for a nine-year-old bilingual child would lie above continuum 1. Such a schema could then be used to plot the D scores for bilingual children of various ages and thus determine the role each of their languages has to play in their CS utterances in terms of the ML/EL asymmetry.

As was the case with the Word Frequency schema, there is no reason why this Vocabulary Diversity schema could not be used to interpret D scores of tri- and multilingual children: the score for each language plotted on an appropriately aged continuum would reflect the degree to which each language is acting as a ML or an EL. The schema could also be extended to be used with teenage and adult data - additional Vocabulary Diversity continua could be established for more accurate

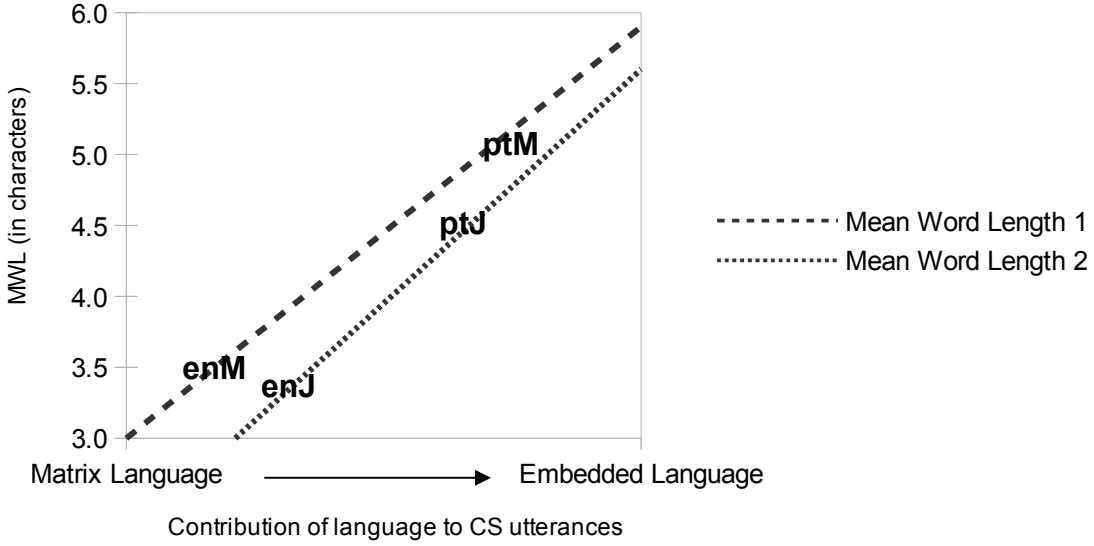
interpretation of the D scores of such speakers. In the case of the adult bilingual speakers in the LOBILL Corpus, the D scores for addressee-specific CS utterances were either unavailable (as in the case of PAI whose code-switching did not offer enough tokens) or significantly skewed by idiosyncratic language use (recall the case of 'olha' in MOT's code-switched utterances (4.2.2.2)). Therefore, I am unable to propose a predictive continuum for bilingual adults.

It is evident that the proposed schema is a preliminary attempt at representing the relationship between vocabulary diversity and the roles of the languages participating in code-switched speech. Its application to other bi- and multilingual CS data sets would allow for further exploration of this relationship which would then, in turn, allow for further refinement of the schema. Such is the case of the fourth and final relationship, which is explained in more detail in the following section.

8.1.3 Interpreting Mean Word Lengths according to the schema

The proposed relationship in (iii) stated that a low mean word length indicates a ML while a high mean word length indicates an EL. However, as was the case with the Vocabulary Diversity relationship, it became clear that any schema had to be able to incorporate developmental aspects if it were to more accurately reflect the roles that each language was performing in the CS utterances of the siblings. The results shown in Fig. 19 (see 4.3.1.1) revealed that the MWL for both English and Portuguese were higher for MEG than for JAM (3.51 and 5.11 as opposed to 3.41 and 4.49). Although both sets of results reflect an ML/EL asymmetry, the plotting of MEG's slightly higher MWL for English on the same continuum as her brother would mean that English was being less Matrix like in her CS utterances than in JAM's. Thus, again, I propose the use of two different continua in order to take into account the increase in word length due to linguistic, developmental differences. These continua are shown below:

Figure 28. Schematic representation of the relative roles of English (en) and Portuguese (pt) in JAM (J) and MEG (M)'s CS utterances addressed to MOT based on Mean Word Length values



As for the Vocabulary Diversity continua, the distance between the MWL continua becomes less as we approach the Embedded Language end. This is a reflection of the frequency of the differential morpheme contribution of the ML/EL to CS utterances which is affected by the developmental differences between the siblings. That is, the high frequency of grammatical morphemes contributed by the ML means that differences in word lengths between the siblings (due to MEG's use of comparatively more complex grammar) would be more in evidence. Any developmental differences between the siblings in terms of their use of content morphemes (the typical contribution of the EL) would be less in evidence due to their lesser frequency.

It is important to point out that in my study preliminary WDLEN analyses had established the comparability of English and Portuguese in terms of word lengths and therefore allowed me to use such measurements to propose the above schema which correlates Mean Word Lengths with the relative roles of languages in CS utterances. For other language pairs the lack of comparability would make the use of my schema problematic. For example the mean word lengths of German compound nouns (commonly written with no intervening spaces) would be longer than their English counterparts (more commonly written with spaces). Of course, during the process of transcription, researchers could make use of conventions (such as the symbol '+' placed between the separate components of the compound) which would

then allow such compounds to be automatically analysed in terms of their individual morphemes. Such decisions, however, would be determined by the research aims for which the corpus is built. Despite this caveat, however, I believe the proposed schema offers an original, empirical, way of establishing the relative contribution of each of a bilingual's (or multilingual's) languages in their code-switched utterances in terms of the ML/EL asymmetry.

In these last three sub-sections (8.1.1, 8.1.2 and 8.1.3) I have sought to explain in more concrete terms how four types of quantitative measures can be used to characaterise a speakers' asymmetrical language use in CS speech. Although the proposed schema is based on the results of analyses using particular CLAN commands (FREQ, VOCD and WDLEN), this does not mean than its application is restricted to the analysis of corpora transcribed according to CHAT conventions. It is likely that the output of other software programmes used to measure word frequency and word and utterances length would provide similar results which can still be interpreted according to the schemas in 8.1.1 and 8.1.3. And with regards to vocabulary diversity scores, it is feasible to assume that one could simply replace the D-score scale (found on the y-axis of my schema) with that of an equally valid scale of measurement. Although it may be beyond the remit of this study to investigate the wider applicability of the schemas presented above, it is evident that there is potential for them to be used by researchers examining bilingual corpora data quantitatively.

8.2 Theoretical contributions of a word and code-level investigation of code-switching

In this section, I will comment on the results of my word and code-level analyses of code-switching in the LOBILL Corpus in terms of how they contribute to our current understanding of both the grammatical and pragmatic nature of this bilingual phenomena. This will involve summarising what my word frequency results reveal about the differential grammatical contribution of the languages participating in code-switched utterances (8.2.1) and highlighting the relationships found between code-switching and linguistic phenonema such as retracings and reformulations, errors, tag questions and metalinguistic usage (8.2.2).

8.2.1 Word frequency results and the 4-M Model

As described in 2.2.2.1, Myers-Scotton's 4-M Model offers a classification of morpheme types based on whether they are conceptually activated or structurally assigned: in Classic code-switching the Matrix Language typically contributes the morphemes which make up the grammatical framework of the utterance (such as early system, bridge and outsider morphemes) while the Embedded Language typically contributes content morphemes, examples of which are nouns, verbs, and adjectives. Due to the language and addressee coding in the LOBILL Corpus I was able to use FREQ to produce word lists for each language per speaker/addressee which I was then able to analyse in terms of different morpheme types and interpret according to the ML/EL asymmetry.

A detailed examination of the morpheme types found in the siblings' word lists (top 20 occurrences) from Meal Time and Telephone Interactions (see 5.1.2 and 5.1.3) provided clear evidence of an ML/EL asymmetry at work in their code-switched utterances when addressing their mother and father. When the siblings addressed their mother in meal time interactions, English played the role of Matrix Language while Portuguese contributed as the Embedded Language; when they addressed their father over the telephone the reverse was found to be true. A comparison of JAM and MEG's lists revealed that their use of the Matrix Language could be considered similar in nature: when addressing MOT 12 out of the top 20 occurrences for English appeared in both lists; when addressing PAI the number of shared words in the Portuguese word lists was a substantial 15 out of 20.

A comparison of the frequency of content words in JAM and MEG's English and Portuguese word lists addressed to MOT in all of the interaction types (5.1.3) again revealed an asymmetry in terms of the ML/EL. This was seen as evidence to support the claim that the siblings' language use was more dependent on the interlocutor variable than on the nature of the interaction. Slight differences between the siblings in the frequencies and types of content words occurring in their English and Portuguese lists were interpreted as further evidence that MEG's use of both the Matrix and Embedded Language was more classic than JAM's (see discussion of Table 13 in 5.1.3).

The analysis of the word lists for MOT (5.1.4) revealed a very strong asymmetrical pattern in terms of the morpheme types contributed by each language: her use of English as the ML and Portuguese as the EL when addressing her children was interpreted as being very classic. This contrasted with what the siblings'

word lists addressed to each other appeared to indicate - that neither English nor Portuguese appeared to be assuming a definite ML or EL role in their code-switching with each other. As we saw in Chapter 7, the utterance-level analyses revealed that it was indeed possible to identify an asymmetry in their use of English and Portuguese with each other, indicating that the word-level results were not an accurate reflection of the true nature of the code-switching occurring between the siblings. The reason for such discrepancies was attributed to methodological issues related to the inclusion/exclusion of multi-addressed CS utterances. These issues and a resolution to the problem are discussed in 8.4.

Also noted from the examination of the word lists was the significant presence of pairs of translation equivalents such as the/a,o, and/e, I/eu, no, isn't, don't/não, yes/é, to/para, that,which/que and look/olha. Although the frequent occurrence of the same words in both language lists would not be predicted by the 4-M Model, the utterance-level analysis of the informants' CS utterances revealed that they mostly appeared together, the result of a speaker retracing and reformulating after an involuntary usage of an EL item. Thus there was a logical explanation for the appearance of these translation equivalents in the lists and they did not actually constitute a violation of the differential morpheme principle on which the 4-M Model is based.

Although Myers-Scotton's MFL Model and accompanying 4-M Model have come under criticism from those who align with the 'no special constraints' theory proposed by MacSwan's Minimalist Programme, the fact that they have been used as a theoretical framework in so many studies of code-switching does appear to prove that they are of significant value[197]. Indeed, when selecting a Model for their analysis of corpus-based CS clauses, Carter et al (2011) found the MFL Model to be the only one to meet their three criteria[198], pointing out that the competence-based MP model was not designed to deal with production data. And judging by MacSwan's paper presentation at a recent (2013) conference on code-switching[199], it appears that no attempt is being made to validate the MP model via naturalistic corpus data: his data base for analysis consisted of 50 questionnaires designed to elicit grammaticality judgements! It would be interesting to see how he would deal with the code-switched

[197] Selected studies include Lanza (1997), Vihman (1998), Paradis et al (2000), Pittman (2008), Liu (2008) and Carter et al (2011).
[198] The three criteria were the following: (I) Designed to deal with production data, (ii) Can analyze individual clauses and (iii) Applies to both monolingual and bilingual clauses. (Carter et al, 2011:157).
[199] "Constraint-free Code-switching and DP-internal Word Order"(MacSwan, 2013).

data in the LOBILL Corpus and how he would account for the CS patterns my analysis has revealed. As it is, the quantitative and word-level results of my study have provided ample evidence to support the main principle underlying the MFL Model which is that code-switching is characterised by an asymmetry in terms of how each language participates in bilingual utterances.

While the word-level analyses of the CS data in the LOBILL Corpus were concerned with uncovering the relationships between morpheme types and the ML/EL, the code-level analyses allowed me to investigate other types of relationships, as will be seen below.

8.2.2 The contribution of code-level analyses to the investigation of code-switching in naturalistic data.

In the second half of Chapter 5 (5.2), I showed what a frequency analysis of five different codes could reveal about different aspects of the code-switching practice of the bilingual speakers in the corpus. For four of these codes the focus was on comparing the frequencies of the coded phenomena in the informants' CS utterances to that of their frequencies in monolingual utterances. For the remaining code, the specially designed CS postcode, the interest was in investigating and comparing the frequencies of the different combinations of letters (e for English and p for Portuguese) found in the postcodes at the end of each sibling's CS utterances (see 5.2.1).

The examination of the CS postcodes was very fruitful in being able to shed light on similarities and differences between the siblings in terms of how English and Portuguese patterned in their CS utterances. By incorporating the variable of addressee into the analyses, I was able to show that the siblings' use of English-initiated and Portuguese-initiated utterances again reflected the ML/EL asymmetry found elsewhere: when addressing their mother, the percentages of postcodes beginning with 'e' for JAM and MEG were 77% and 87% respectively while those beginning with 'p' addressed to their father accounted for 81% and 84% respectively. This evidence was seen as lending support to the idea that the language in which a speaker initiates an utterance would be a good indicator of the role of that language in a CS utterance i.e as the ML. In quantitative terms, high percentages would correlate with the Matrix Language while relatively low percentages would indicate the Embedded Language.

A comparison of the different CS postcode variants in JAM and MEG's frequency lists (see Tables 15 and 16) revealed that their code-switching patterns were very similar too, the most frequent variants being the same and accounting for the vast majority of the data (when addressing MOT, 6 variants accounted for 93% and 96% of JAM and MEG's total number of postcodes and when addressing PAI, 4 variants accounted for 84% and 90%). However, this comparison also revealed that JAM made use of more CS variants than MEG (twice as many), some of these variants involving several switches back and forth. Apart from showing more variation in his CS patterns than MEG, the overall higher frequency of his CS postcodes when compared to MEG was yet more indication of the fact that JAM was a more prolific code-switcher than his sister.

It is important to remember that all of the above observations were gleaned from the interpretation of frequency lists of postcodes and not from an analysis of the utterances themselves. While this highlights how insightful such a methodological approach can be, it is evident that qualitative analyses would be necessary to enable the confirmation of such observations. This turned out to be especially so when I analysed the frequency postcode lists of the siblings resulting from their code-switching with each other (Table 17). The finding that there was a more balanced proportion of English-initiated and Portuguese-initiated utterances provided evidence for symmetrical, rather than asymmetrical language use between the siblings. Qualitative analyses (see 7.2) revealed that this was not the case and that a methodological problem with the selected data had skewed the results. This is discussed fully in 8.4.

As mentioned in the introductory part of this section, the examination of the other four codes involved comparing code-switched data with monolingual data. In 5.2.2, my comparison of the frequencies of the codes for retracing and reformulations ([//] and [///]), led me to conclude that for both JAM and MEG (but more so for JAM), retracings and reformulations were a significant feature of their CS utterances. I suggested that it was likely that the frequency of retracings and reformulations must, in part, be due to the siblings' switching to the other language in order to effect a repair after involuntary usage or in order to better express themselves, semantically and/or syntactically. The slightly higher frequencies for JAM when compared to MEG were interpreted as reflecting his need to retrace and reformulate more due to his comparatively less developed command of his two languages. My analysis ended

with the observation that it was possible to posit a relationship between the results of my Mean Utterance Length analyses and the results of my frequency analyses of the retracing and reformulating codes: the higher MULs of CS utterances could be attributed, in part, to the higher incidence of retracings and reformulations occurring in bilingual utterances.

In my analysis of the error codes occurring in CS and monolingual utterances (see 5.2.3), the frequency results showed that both children produced slightly more errors in CS utterances than in monolingual utterances. This was considered to be unsurprising given the potential conflicts arising from combining two different languages in a single utterance. It was the analysis of the words coded as errors (see Table 22) which proved to be more insightful. For both JAM and MEG, the majority of the top 20 most frequent errors involved English words (15 and 14 respectively). This might be interpreted as reflecting the fact that the siblings' English was less developed than their Portuguese (the dominant language of the community environment). A comparison of the top 20 morpheme types found in each list hinted at a difference between the siblings in the nature of their errors: while most of MEG's errors involved content words, the majority of JAM's were the result of erroneous use of grammatical morphemes. It was posited that with both language systems less developed than MEG, JAM would be more susceptible to grammatical errors when code-switching[200]. More evidence to support this supposition was found when I examined the errors in their linguistic context (discussed in 8.3).

In 5.2.4, the evidence provided by the frequency analyses of the code [@tq] pointed to clear differences in tag question usage between the bilingual speakers in both CS and non-CS utterances. Despite being a prolific user of tag questions with 375 occurrences overall, MOT's use of them in CS utterances was virtually non-existent (restricted to only the two tag questions 'yeah?' and 'is it?'). In MEG's data there were also only 2 (out of 48) that occurred in her CS utterances (both 'yeah?'). This contrasted noticeably with JAM's data where 26.8% of his tag questions (32 out of 119) were used when he was in bilingual mode. A comparison of the types of tags occurring in CS and non-CS data revealed that JAM showed a clear preference for two particular types when in bilingual mode, the invariant 'yeah?' and 'isn't it?'. The longitudinal analysis of JAM's tag question usage carried out in 6.4.2 suggested that

[200] It is important to note that this does not mean that 'content' errors were less frequent for JAM than for MEG. It just means that JAM's content errors were further down his word lists.

developmental factors were responsible for the significant differences in the results between the siblings in terms of the overall frequencies of tag questions and in terms of their occurrence in CS data as opposed to non-CS data. This will be explored more fully in 8.3.

The importance of considering other influencing factors when interpreting CS data became more apparent when I carried out the analysis of the metalinguistic code (['']) in 5.2.5. As far as JAM's results were concerned, the frequency analysis revealed that he engaged in more metalinguistic language use while in monolingual rather than bilingual mode, and that when code-switching, his metalinguistic references mostly involved single words. MEG's results showed that she made relatively more use of this language device when code-switching compared to when she was speaking in monolingual English or Portuguese. It also showed that, compared to JAM, her metalinguistic usage when in bilingual mode often involved the use of more than just single words. It was the analysis of MOT's results which proved to be most telling. A comparison of her metalinguistic usage in CS and non-CS data revealed that such usage was a particularly significant feature of her code-switching practice: 12% of the codes occurred in only 1.8% of her total token count (her CS tokens). And on examining the word types coded with [''], I discovered a further difference between the siblings and their mother. While there was a roughly equal balance of English and Portuguese words in the siblings' lists, the overwhelming majority of MOT's words coded with [''] were Portuguese. Cross-referencing with other data allowed me to discover that these Portuguese tokens actually represented over 25% of all of the Portuguese tokens used by MOT when code-switching. This finding confirmed that one of the particularly important functions of the Embedded Language (Portuguese) in MOT's code-switched utterances was to refer to language metalinguistically. And despite PAI's lack of data, considering that 23.8% of his CS tokens (which represent 4.7% of his total tokens) were coded with [''], these results serve to highlight a potential difference between the siblings' and parents' code-switching practice in terms of function. These differences in metalinguistic usage were further examined in 6.5 and are summarised in section 8.3 below.

The five types of code-level analyses carried out on the LOBILL Corpus provided insights into various aspects of the bilingual speakers code-switching practice. Patterns in the data were detected and comparisons between the speakers allowed differences to be highlighted. Factors affecting the differences in occurrences

were considered and potential relationships were suggested between the frequencies of the coded phenomena and both grammatical and functional aspects of code-switching. Such insights were only made possible due to the particular methodology used in my study.

8.3 The contribution of utterance-level analyses to the investigation of the siblings' code-switching practices with their parents

In Chapter 6 I examined a selection of CS utterances in order to search for explanations for certain findings which arose out of the quantitative and word- and code-level analyses carried out and discussed in Chapters 4 and 5. It became evident that such examination often necessitated a detailed analysis of the original transcription as a straightforward surface-level analysis of some of the CS utterances would not suffice. As will be highlighted in this section, my qualitative examination proved to be very productive and shed light on various aspects of the nature of the siblings' code-switching, including the motivations underlying its use.

The purpose of the utterance-level analyses carried out in section 6.1 was to search for explanations as to why JAM and MEG should inititate a minority of their CS utterances in the Embedded Language (Portuguese when addressing MOT and English when addressing PAI). Through the discussion of several examples it was possible to highlight the motivations behind the occurrence of these exceptions to the code-switching patterns found in the majority of the data. For both siblings, these motivations included conscious ones, such as switching languages in order to quote somebody or refer to something metalinguistically, but involuntary switches were also seen to account for a significant number of the exceptions. In some cases, changes in the linguistic environment (due to holidays in England and the siblings' eventual move there) also played a part in influencing JAM and MEG's use of the EL to initiate CS utterances. It became evident that JAM was more susceptible to such changes while MEG's more developed language awareness ensured that she was more in control of how her EL contributed to her code-switching.

Differences between the siblings in terms of language development and awareness were held responsible for the differences found in the analysis of the retracings and reformulations employed by the siblings in their CS utterances (see 6.2). Although the examples revealed that the siblings had recourse to this strategy for similar purposes, that is, to accomodate to their interlocutors' linguistic

preferences after a mostly involuntary use of the EL, their degree of success was markedly different. MEG's retracings and reformulatings into English or Portuguese showed evidence of proficient self-monitoring and effective self-repair as she strove for linguistic consistency with each interlocutor. JAM's use of this strategy, however, clearly showed that he found it more of a challenge to maintain this type of consistency, especially with his mother: this was evidenced by his more frequent need to revert to Portuguese in order to be able to express himself, especially in emotional situations. That JAM should fall back on Portuguese was clear indication that the latter was for him the more dominant language of the two.

Another important finding from the analysis of the CS utterances containing retracings and reformulations was that relating to the occurrence of translation equivalents, originally detected in the frequency word lists. It became clear that their occurrence did not constitute contra-evidence to the 4-M Model at all but merely reflected the siblings' use of retracings after involuntary usage.

The longitudinal utterance-level analysis of the siblings' errors in CS utterances (see 6.3) confirmed the interpretations made at the code-level analyses of their errors (see 5.3). Whereas MEG's errors were fewer and mostly lexical in nature, JAM produced more than three times as many errors, the most frequent being grammatical in nature. The qualitative analysis further revealed that for both siblings the influence of Portuguese accounted for the majority of their CS errors, this influence involving both grammatical and sociocultural linguistic transference from Portuguese into English. However, my longitudinal analysis enabled me to show how such influence was susceptible to changes in the linguistic environment and how the increasing linguistic competence of the siblings (particularly in JAM's case) was ultimately responsible for the notable decrease in errors in CS utterances occurring over the last periods of data.

Both the code frequency analyses and utterance-level analyses highlighted the fact that slightly more errors tended to occur when the siblings were engaged in code-switching (rather than speaking monolingually). However, the discussion of one particular example (104) brought to the fore how code-switching also represented a strategy whereby a speaker could avoid potential errors resulting from grammatical and/or lexical transference from one language on to another. Of course, this use of code-switching as a communicative facilitator would only be possible in interactions with bilingual addressees.

Due to the differences in complexity of the tag question systems in Portuguese and English, the longitudinal analysis of JAM's use of tag questions in both mono and bilingual utterances (see 6.4.2) proved to be particularly revealing. There was ample evidence to conclude that JAM's ubiquitous, and mostly erroneous use of the generic 'isn't it', in both his mono and bilingual utterances was the result of direct transference of the Portuguese generic tag 'né'. By tracking the development of his tag question usage over three years it was possible to see that only after the age of six did JAM really begin to grasp the complexity of the English canonical tag question system. Intensive exposure (while on holiday and after his move to England) resulted in increased accuracy and more varied tag question usage but this was restricted to his monolingual utterances – JAM ceased to use tag questions at all in his CS utterances. It was posited that his increasing language competence and awareness resulted in his 'playing safe' in bilingual utterances by avoiding the use of tag questions. The use of such an avoidance strategy was given as the reason why tag questions were virtually non-existent in the CS data for his older sister MEG.

In contrast to the findings above, where there was an inverse association between language competence/awareness and the occurrence of tag questions in CS utterances, the utterance-level analysis of the metalinguistic codes ['] in 6.5 revealed a positive relationship: increased linguistic competence tallied with more frequent occurrence of metalinguistic codes. MOT was the most frequent user of this function of code-switching, followed by MEG and then JAM. While many of the siblings' codes were simply marking the quoting of a variety of speakers who were not present in the interactions, the majority of MOT's codes were seen to mark the quoting of her own children's words. Much of this quoting had the function of directing the siblings' attention to their use of a particular word or expression in order to elicit a response, such as a translation into English. Metalinguistic discussions between MOT and her children were the loci for the occurrence of many more of the metalinguistic codes and the analysis of these particular utterances together with additional analyses carried out on the use of language labels ('Portuguese', 'English', 'português' and 'inglês') allowed for comparisons to be made between the siblings in terms of how they used and understood their two languages metalinguistically. And although the evidence from the longitudinal analysis of the language labels supported the idea that the differences between JAM and MEG were related to linguistic maturation, it also pointed to the influence that contextual factors, such as the

linguistic environment, could have in affecting the development of a bilingual child's language awareness.

It is perhaps important to point out that all the insights afforded by my utterance-level analyses described above have resulted from the examination of only a selection of the CS data in the LOBILL Corpus. This selection was ultimately determined by the coding in the corpus which I wished to exploit to its full potential. My methodological approach has allowed me to carry out a systematic and detailed analysis of different batches of CS utterances and compare them to equivalent batches of monolingual utterances. By combining a frequency-based approach with a qualitative analysis I have been able to make comparisons between the bilingual speakers which, in turn, have allowed me to propose relationships between code-switching and the occurrence of certain linguistic phenomena (retracings and reformulations, errors, tag questions and metalinguistic usage). Due to the heterogenous and longitudinal nature of the data my analyses have been particularly insightful as I have been able to consider the influence of contextual factors and differences in linguistic development on the occurrence of the phenomena being examined.

8.4 The contribution of the analyses of the parents' code-switching and of that occurring between the siblings

Although it was the interactions between the siblings and their parents which provided most of the data for this study, the analysis of the code-switching occurring in 8 other speaker/interlocutor combinations (see Chapter 7) proved useful in several aspects. Apart from enabling me to examine parental code-switching practices and consider their influence on the siblings' practice, it also allowed me to ascertain the reasons for the unexpected findings from quantitative analyses that indicated that there was no identifiable Matrix Language in the code-switching occurring between the siblings.

With regards to MOT's and PAI's code-switching with their children, the utterance-level analyses revealed that their code-switching was typically classic in style with their mother tongue providing the morphosyntactic frame and their second language being used very sparingly as the Embedded Language. For MOT, the majority of the Portuguese EL insertions constituted single items which were either quoted words, metalinguistic references or content words intrinsically related to the

family's shared sociolinguistic and cultural environment. PAI's insertions in English served the very specific metalinguistic purpose of drawing his children's attention to their own code-switching practices in order to encourage more appropriate usage.

In terms of the code-switching occurring between PAI and MOT, only 6 CS utterances each were available for analysis. Despite this reduced number, in PAI's 6 utterances it was possible to identify the same asymmetrical language use found above, this time the contribution of the English EL items being entirely lexical in nature. It was the analysis of MOT's 6 CS utterances addressed to PAI that proved to be most revealing. Whereas 3 of the utterances were structured as expected, with English as the ML and Portuguese as the EL, in the remaining 3 there was a reversal of roles. This reversal was interpreted as being a deliberate attempt by MOT to counter the drastic reduction in the siblings' exposure to Portuguese which came about as a result of the family's move to England. This interpretation was supported by further frequency analyses which showed that while PAI's use of English with his children and wife increased after the move, MOT's use of English actually decreased, and this despite the potential influence of the immediate linguistic environment.

The analysis of MOT and PAI's CS utterances highlighted one important aspect of their potential roles as agents of language socialization, that of paying overt attention to their children's code-switching practices. However, while MEG with her comparatively heightened language awareness appeared to respond to such attention, JAM appeared to pay less notice to his parents' metalinguistic comments. The fact that MEG's code-switching practice was found to be of the more classic type and similar to her parents' bilingual usage, may suggest that parents also have indirect roles to play in terms of modeling bilingual language use. However, the fact that JAM's code-switching practice diverged somewhat from his parents' practice implies that other factors may exert a stronger influence, especially on a child whose two languages are still developing and thus more susceptible to external influences.

The potential influence of the linguistic environment was key to interpreting a minority of the code-switched utterances exchanged between the siblings (see 7.2). In the majority of both JAM and MEG's CS utterances, asymmetrical language use was easily identifiable with Portuguese in the role of the ML and English taking on the more limited role of the Embedded Language (mostly contributing single word items). However, there were exceptions to this pattern and explanations were only forthcoming following a thorough examination of the linguistic context in which each

one occurred. While in MEG's case such examination of the discourse revealed more about her language awareness (see discussion of example (154)) and showed her purposeful use of code-switching (see discussion of example 157), in JAM's case one could attribute his exceptions to the influence of the immediate linguistic environment: his typical CS pattern when addressing his sister changed after he moved to England, English taking on a more dominant role. Due to the lack of CS data for MEG, this influence is less evident in her bilingual interactions. However, the results of word frequency analyses which compared the siblings' use of English and Portuguese with each other before and after the move, reflected a change in language dominance in their interactions. In less than 4 months, English appeared to have replaced Portuguese as the more normal form of communication between JAM and MEG.

As mentioned before, the finding that there was a clear asymmetry at work in the siblings' CS utterances addressed to each other was contrary to my interpretation of the quantitative analyses carried out and discussed in Chapters 4 and 5. There the results had indicated a more balanced participation of English and Portuguese in their CS utterances. What came to light as a result of my utterance-level analyses was the importance of being able to exclude multi-addressed utterances from quantitative analyses. This methodological insight, first raised in Chapter 7 is fully explored in the next section.

8.5 Methodological issues: including and excluding addressees

In the last chapter, a potentially important methodological issue was raised regarding the inclusion of multi-addressed utterances in CLAN analyses (see 7.2). It became evident that it would be important to be able to exclude multi-addressed utterances from those analyses where specific speaker/interlocutor combinations were under scrutiny. Whereas for utterance-level analyses this exclusion could be done manually (by simply ignoring those utterances in the output addressed to more than one person), this was not possible for quantitative analyses. In this section I will present my solution to this problem, thereby offering a method for others who face the same methodological issue.

In order to carry out each of the addresse-specific analyses in my study, my method was to use KWAL to select those speaker utterances addressed to a certain interlocutor. For example, the following command line was used to output a

frequency list of English words found in the CS utterances JAM addressed to his father (see 4.1.4 for further details):

kwal @ +t%add +t*JAM +s"PAI" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5

By changing the string -s"<@pt>" to -s"<@en>" the output would then be a frequency list of Portuguese words found in the same CS utterances. After carrying out the utterance-level analyses (Chapters 6 and 7) it became evident that KWAL selected any utterances where PAI occurred on the addressee tier, including those where other addressees were present. What was needed was a way of excluding these multi-addressed utterances from the analysis.

One way of excluding such material is to remove the string +s"PAI" and instead use the string -s"*speakercode*" for each unwanted addressee (18 in the case of the LOBILL Corpus). By doing this, KWAL would exclude any utterances addressed to those speakers and by default include only those addressed to the speaker whose code does not appear in the command line (i.e. PAI). With such a method, the above command line would now look like this:

kwal @ +t%add +t*JAM -s"MOT" -s"MEG" -s"JAM" -s"BEC" -s"GRA" -s"GRD" -s"JAK" -s"MAX" -s"WIL -s"ARL" -s"AVO" -s"DAN" -s"JAN" -s"JUL" -s"ROS" -s"SAR" -s"VIN" -s"VOV" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5

In order to focus on a different addressee, for example, MOT, one would just need to remove "MOT" and replace it with the missing code, "PAI". This way KWAL would now select only JAM's CS utterances addressed solely to his mother.

Although the method above would ensure effective retrieval of the desired input, the command line is very long. This does not present a problem in terms of typing in the CLAN command window as one can recall a previous command line and just make a simple substitution (e.g "PAI" for "MOT"). However, it is possible to reduce the size of this command line by using another of CLAN's facilities, an exclusion file, an example of which can be seen in the command line above: -s"@nonwords.cut". This particular exclusion file (nonwords.cut) contains a list of nonwords (such as 'err' and 'mmm' etc) which I wanted to exclude from all frequency

analyses. By using this string, CLAN refers to the file nonwords.cut (stored in the LIB) and ensures that any words in that file are automatically excluded from the output.

Through experimentation I discovered that such a method could also be applied to speaker codes: any three-letter speaker code appearing in an exclusion file would mean that any utterances addressed to that speaker would be excluded from the input by KWAL. Due to the nature of my particular corpus and research purposes, I decided to create an exclusion file named monoadds.cut (where monoadds stands for 'monolingual addressees') which contains a list of all the 15 speaker codes of the monolingual speakers in the LOBILL Corpus. This way, instead of cluttering the commands window with 15 individual speaker codes, I could group them all under one heading, the command line now looking like this:

kwal @ +t%add +t*JAM -s"MOT" -s"MEG" -s"JAM" -s"@monoadds.cut" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5

It was necessary to maintain the individual bilingual speaker codes in the command line as without them I would not be able to effect the necessary substitutions for each different analysis.

Thus, through experimentation and testing I have arrived at a solution to my methodological problem involving multi-addressed utterances. Clearly, such insights would have been much more useful had they occurred earlier on in my study as within the time contraints of my research it is not feasible to repeat all of the pertinent frequency analyses. However, I do feel it is important to demonstrate the extent of the effect of the inclusion of multi-addressed CS utterances in such analyses and I will do this my comparing some original frequency results with those obtained by the method discussed above. This way it will be possible to consider the validity of my original results in the face of this methodological issue.

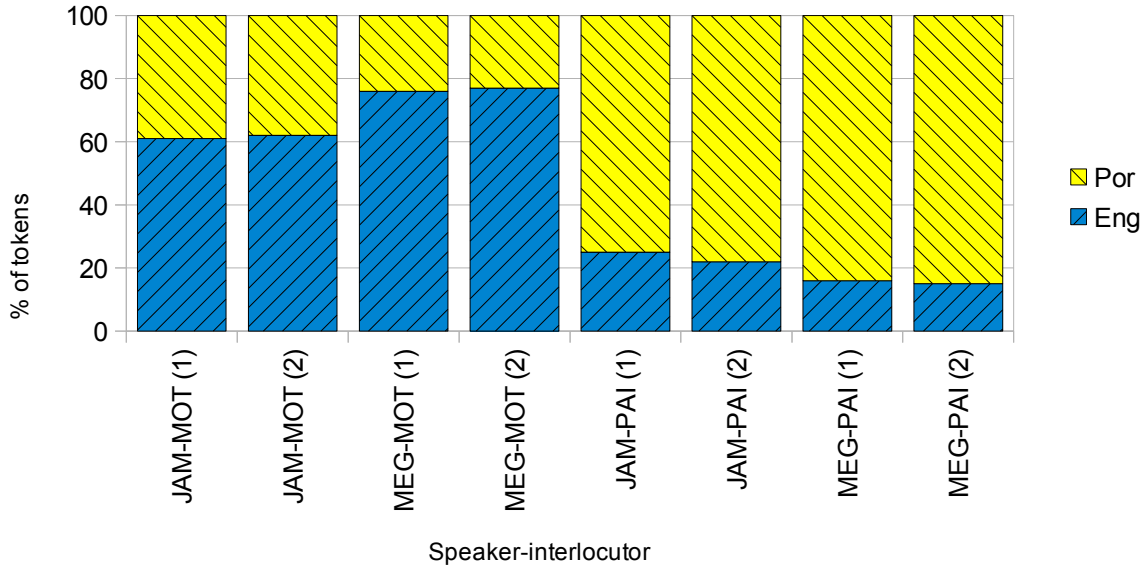
8.5.1 The effect of the exclusion of multi-addressed CS utterances on my results

In Fig. 7 (see 4.1.4) we have the frequency results of the numbers of English and Portuguese tokens in CS utterances for the 12 speaker/interlocutor combinations involving the four bilingual informants. The input data for these results included any multi-addressed CS utterances. I decided to take 6 of the speaker/interlocutor combinations (JAM/MOT, MEG/MOT, JAM/PAI, MEG/PAI, JAM/MEG and MEG/JAM)

and repeat the same frequency analyses, the only difference being that this time any multi-addressed CS utterances were automatically excluded from the input[201]. As I am interested here in comparing the proportion of English tokens to Portuguese tokens, as opposed to raw numbers, I have presented the results in terms of percentages - this will allow for more effective comparison across the data. While Fig. 29 shows the comparative results for the first four combinations (the siblings addressing their parents), Fig. 30 shows those for the remaining two combinations (the siblings addressing each other). In both cases the original results (1) are shown immediately next to the new results (2).

Looking first at Fig. 29 it does appear that the difference between both sets of results is almost negligible. For the first speaker/interlocutor combination shown (JAM-MOT), there is a 1% increase in English tokens (from 61% to 62%) and a correspnding 1% decrease in Portuguese tokens (from 39% to 38%). Exactly the same increase/decrease is seen for MEG-MOT, the percentage of her English tokens going from 76% to 77%.

Figure 29. Proportions of English and Portuguese tokens in the siblings' CS utterances addressed to their parents including (1) and excluding (2) multi-addressed utterances

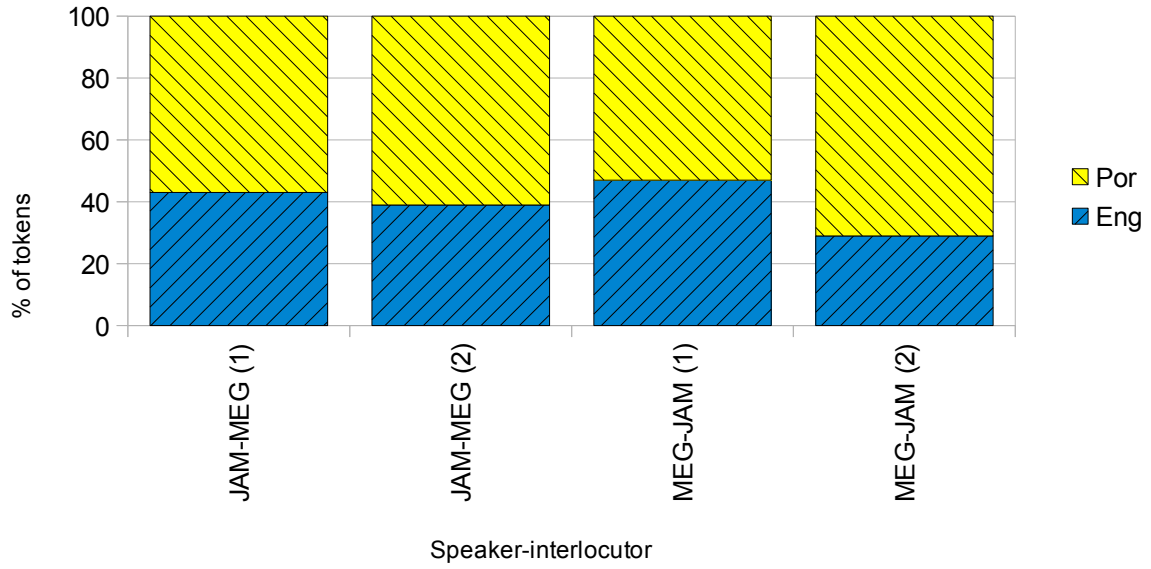


[201] The following two analyses were repeated for each speaker/interlocutor combination: kwal @ +t%add +t*JAM -s"MOT" -s"MEG" -s"JAM" -s"@monoadds.cut" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@pt>" +r5 ; kwal @ +t%add +t*JAM -s"MOT" -s"MEG" -s"JAM" -s"@monoadds.cut" +u +d | freq +o +s"[+ *]" -s"@nonwords.cut" -s"<@en>" +r5

With regards to PAI as the addressee, we find a slight increase in the percentages of Portuguese tokens in (2): for JAM this increase is from 75% to 78% and for MEG it is from 84% to 85%. Although such increases are small, these findings do show that the inclusion of multi-addressed CS utterances in frequency analyses needs careful consideration. In the case of my results, the slight increase shown in the siblings' use of the interlocutor's mother tongue (English for MOT and Portuguese for PAI) actually reveals that the Matrix/Embedded Language asymmetry originally indentified in the code-switching patterns of the siblings when addressing their parents is even more asymmetric, or 'classic' than previously thought.

These results clearly confirm the effect an additional interlocutor can have on a speaker's code-switching practice and show how important it is to be able to account for this factor when interpreting quantitative results. The influence of this multi-addressee variable is more evident in the contrastive results shown below in Fig. 30, especially as far as MEG is concerned.

Figure 30. Proportions of English and Portuguese tokens in the siblings' CS utterances addressed to each other including (1) and excluding (2) multi-addressed utterances



If we contrast the results for JAM's CS utterances when addressing his sister (first two columns), we find a 4% increase in Portuguese tokens (from 57% to 61%) when the multi-addressed CS utterances are excluded from the input. For MEG (when

addressing her brother) the increase in Portuguese tokens amounts to 17% (from 53% to 71%). If we interpret these new frequency results (2) in terms of what they indicate about the existence of an ML/EL asymmetry, what we find is that for MEG there is now a significant contrast between the two languages: with 71% of CS tokens being contributed by Portuguese and only 29% by English, Portuguese does appear to be in the role of the Matrix Language. These results support what was found when I examined MEG's utterances qualitatively (see 7.2.2). The results for JAM do not indicate such contrast in terms of his asymmetrical use of the two languages: 39% of the CS tokens are contributed by English and 61% by Portuguese. However, these percentages are still more disparate than the previous results in (1) where the relative percentages were 43% and 57%. The new results more accurately reflect the extent of the asymmetrical language use found in JAM's code-switching with his sister (see 7.2.1).

It is clear that the exclusion of multi-addressed utterances has had a greater impact on the frequency results of the siblings interactions (when addressing each other) than on those where they code-switch with their parents. It would be wrong to assume that this is simply because more of the CS utterances exchanged by the siblings (when compared to those addressed to their parents) were actually multi-addressed. It might be that accommodation to the language of the additional addressee(s) has had an effect on the results. However, a look at the data in the table below (particularly column 4) reveals that in this case, multi-addressed CS utterances were indeed much more frequent in the exchanges between the siblings.

Table 28. Totals of CS tokens including (1) and excluding (2) multi-addressed utterances

Speaker - interlocutor	Original totals of CS tokens (1)	New CS totals (2)	Multi-addressed CS Tokens excluded (% of original total)
JAM-MOT	3,655	3,419	236 (6.4%)
MEG-MOT	2,039	1,895	144 (7%)
JAM-PAI	920	872	16 (3.5%)
MEG-PAI	1,251	1,218	33 (2.6%)
JAM-MEG	280	148	132 (47%)
MEG-JAM	197	68	129 (65.4%)

Although the differences in total tokens across the speakers are not comparable, the percentages are quite revealing: for both JAM and MEG only approximately 7% of the CS tokens addressed to MOT were multi-addressed and less than 4% of those addressed to PAI involved another addressee. These percentages are significantly lower than those shown for the combinations JAM-MEG and MEG-JAM: 47% of the CS tokens JAM addresses to his sister are also addressed to someone else and 65% of the CS tokens MEG addresses to her brother involve another addressee. The removal of these multi-addressed tokens (approximately half of the data in the case of the last two combinations) is understandably going to have an impact on the results. However, it is only by effectuating this removal that it will be possible to reveal a more accurate picture of how both languages actually contribute to the structuring of a speaker's code-switched speech with a particular interlocutor.

The variation in total numbers of CS tokens for the six combinations shown in Table 28 is also worth drawing attention to. If there had been more CS data available for some of the combinations, it is likely that the removal of any multi-addressed utterances would have had less of an impact. This highlights the challenges of working with limited data. Performing valid quantitative analyses on limited numbers of tokens is clearly problematic and that is the reason why in my study some speaker-interlocutor combinations were only examined at utterance-level.

The purpose of the discussion in this section has been to highlight an important methodological issue relating to the inclusion and exclusion of multi-addressed utterances in quantitative analyses. In terms of the use of CLAN, a methodological solution has been put forward and shown to be effective in accomplishing the desired exclusion. This method could be applied to all types of analyses whether they be purely quantitative and/or at the word, code or utterance level. The key point here is that it is now possible to more effectively isolate the addressee variable which, as I have demonstrated throughout this study, is of crucial importance when interpreting many aspects of code-switched data. It is hoped that such transparency in terms of dealing with methodological problems will enable others to approach the analysis of their own (multi-addressed) data in a more methodologically sound way, whether that be through the use of CLAN or via another programme.

8.6 The implications of my study for the future of code-switching research.

It is difficult to imagine how I would have been able to make sense of the CS data in the LOBILL Corpus without first approaching it from a quantitative perspective. My frequency-based approach allowed for the detection of patterns, and differences in these patterns, in the CS utterances produced by the bilingual informants. I was able to formulate hypotheses which then served to guide subsequent, more qualitative analyses. As mentioned before, the current study does not attempt to examine or explain every single CS utterance occurring in the corpus. For example, the examination of the siblings' Portuguese-initiated CS utterances addressed to MOT (see 6.1.1) and their English-initiated utterances addressed to PAI (6.1.2) represented only a small proportion of the available CS data for these speaker/interlocutor combinations: for JAM, they amounted to 22% and 18% respectively of his total CS utterance count while for MEG the proportions were even lower, 12% and 14% respectively. Even when discounting the code-based groups of CS utterances, there still remains a significant number of the siblings' CS utterances addressed to the parents which have yet to be analysed[202].

Apart from examining the remaining data from a qualitative perspective, there is the potential for further more detailed grammatical studies of the nature of the code-switching found in the LOBILL Corpus. KWAL and FREQ could be used to search for certain grammatical constructions which could then be examined longitudinally. For example, the potential is there to investigate gender and number agreement in switched noun phrases and compare results with those of other studies (see Endesfelder-Quick, 2013; Eichler et al, 2013 and Liceras, 2013). The study of bilingual compound verbs (see Edwards and Gardner-Chloros, 2007) could also be the focus of an investigation. As shown by my analysis of the siblings' errors, the LOBILL Corpus represents an extremely rich database for the investigation of language transfer (see Treffers-Daller, 2009b), both in terms of lexical transfer (see Jarvis, 2009) and syntactic transfer (Yip & Matthews, 2000). A more detailed investigation of such transfer could contribute to research on the extent to which typological factors determine the outcome of the grammatical and syntactic nature of code-switches: the English/Portuguese data in my corpus could be compared to

[202] By adding up the frequencies of the five types of codes found in the siblings' CS utterances addressed to MOT and PAI we arrive at a total of 711. The total number of CS utterances produced by the siblings when addressing their parents is 977. This means that there are still over 266 CS utterances awaiting analysis. Considering the fact that different codes may appear in a single utterance and/or occur more than once within the same utterance, it is highly likely that the total number of unanalysed CS utterances is significantly higher than 266.

typologically similar language pairs such as English/Spanish (see Lipski, 1985, and Deuchar & Quay, 2000).

The study of prosodic and phonological aspects of the code-switching occurring in the LOBILL Corpus represent yet another potentially rich scource of investigation. Through the CHILDES database researchers have access to both the transcriptions and recordings, thus making such an investigation possible. Researchers interested in word internal switches would benefit from the systematic coding of these occurrences in the LOBILL Corpus (coded with @mf): their retrieval would be instantaneous with the use of the following command line kwal @ +s"*@mf".

Of course, although the above suggestions for further investigation are related to code-switching, in actual fact the LOBILL Corpus provides a rich resource for many avenues of grammatical, pragmatic and sociocultural aspects of linguistic enquiry, whether they be related to bilingual or monolingual usage. For example, the meal time interactions in the corpus would provide a very fertile field for investigating aspects related to Language Socialization (see Blum-Kulka, 2008). Bilingual meal times (where only the parents and siblings are present) could be compared to those meal times where monolingual speakers are also active participants. This comparison might yield insights into how bilingual children are 'socialized' in terms of accommodating to their interlocutors in these sorts of interactions.

Looking beyond the potential for further exploration of the LOBILL Corpus itself, I would also like to highlight how the methodology I have used could be employed in order to exploit other bilingual (code-switched) corpora. With regards to corpora already in CHAT format, I pointed out in 2.2.4.1.2 that much of the bilingual data in CHILDES would benefit from a more consistent and reliable method of language coding in order to maximise exploitability within and across the corpora. Of course it is most often the case that decisions about coding are determined by the research objectives of a particular project and not by considerations related to future exploitability. For example, when I examined the corpora in the FLLOC project (French Learner Language Oral Corpora)[203], I was unable to perform certain quantitative analyses due to the types of language codes that had been used. Although I was able to carry out frequency analyses on the English single word

[203] The FLLOC database comprises nine separate French learner language corpora transcribed in CHAT format. See www.flloc.soton.ac.uk for further details.

insertions in bilingual utterances (mostly coded with @s)[204], I was unable to do the same with those utterances where English played a greater role: these multi-word English insertions or monolingual alternations had been placed within square brackets so that the morphosyntactic tagger would ignore such utterances[205] – they were of no research interest to the FLLOC team. Although with KWAL it was possible to output concordances of the English material placed in square brackets[206], when I attempted to then use FREQ, VOCD and WDLEN to output frequency lists, D-scores or word and utterance length means of this same material[207], the output proved to be problematic: the frequency output consisted of a list of the very same bracketed utterances, the programme clearly unable to break down these larger bracketed units into separate words; the output for VOCD showed that this programme had included the language tags 'eng' (i.e metalinguistic items) as well as the spoken utterances themselves in its D-score calculation; and the zero output for WDLEN indicated that it had been unable to analyse the bracketed material at all. Although these types of analyses may not have been on the research agenda of those responsible for building the corpora, a simple change in the language coding (such as proposed in the current study) would greatly enhance their potential for future exploitability.

Of the nine corpora which comprise the FLLOC database there is one particular corpus, the Young Learners Corpus contributed by Myles and Mitchel, which does indeed appear to offer more foresight in terms of language coding: instead of bracketing off any English material in each utterance they have added either E (for English) or F(for French) to each speaker code and then coded any other-language insertion with the @s method. This coding is illustrated in the example utterances below which were uttered by the speaker INV (Investigator) from the file MT_RP_FM_Y1_Romain_Paulette_Julien (1) [208]:

INVE: a garçon@s:fra$n is a boy.

INVF: so@s:eng$conj <il a>[/] il a trois ans.

---

[204] freq @ +s"*@s*"

[205] Two examples from file 01I19SAR.cha of the Linguistic Development Corpus are coded as follows: *I01: <est ce>[/] est ce que la femme grand [^eng:or] petit and *I01: <elle s'appelle>[/] elle s'appelle[^eng: ouh what am I doing?]

[206] For example, kwal +t*I01 +s"[^*]"

[207] kwal @ +t*I01 +s"[^*]" +d | freq +s"[^*]", Kwal @ +t*I01 +s"[^*]" +d | vocd +s"[^*]", Kwal @ +t*I01 +s"[^ *]" +d | wdlen +s"[^*]"

[208] Files can be downloaded by following the appropriate links on the FLLOC homepage www.flloc.soton.ac.uk

While such coding means that I can now carry out FREQ, VOCD and WDLEN analyses by using simple command lines such as freq @ +t*INVF -s"*@s:eng*" (for a word frequency list of the speaker's French production) or VOCD @ +t*INVE -s"*@s:fra*" (for the D score of the speaker's English production), these results are neccessarily partial and do not represent the speaker's entire production of that language in a particular file or files. This is because by selecting the tiers separately (i.e. only those coded with INVE or INVF), one would automatically be excluding any English/French material occurring in the other set of tiers. It is not feasible to combine both versions of the speaker code in the command line as the input would then consist of both English and French material from which it would only be possible to exclude those insertions coded with @s. By using the language coding suggested in the present study full retrieval of the relevant language material for analyses would be made possible. As all the corpora in FLLOC are already in CHAT format, such changes to their language coding would be straightforward. For example, in the Young Learners Corpus, one could simply insert angled brackets around stretches of material in the same language followed by the appropriate language code, the example utterances shown above being transcribed as follows:

INVE: a[@eng] garçon@s:fra$n <is a boy>[@eng].

INVF: so@s:eng$conj <<il a>[/] il a trois ans>[@fra].

Although it would not be necessary to remove the E and F from the speaker code, their removal would simplify the construction of command lines: instead of having to remember to select both sets of tiers (+t*INVE +t*INVF), one could simply select by the use of +t*INV.

As for the other corpora in FLLOC where English material is excluded by the use of square brackets, the changes would be just as straightforward, as shown in the following examples (see footnote 190 for source of utterances):

Original coding: *I01: <est ce>[/] est ce que la femme grand [^eng:or] petit

LOBILL coding: *I01: <<est ce>[/] est ce que la femme grand>[@fra] or[@eng] petit[@fra]. [+ fef]

Original coding: *I01: <elle s'appelle>[/] elle s'appelle [^eng: ouh what am I doing?]

LOBILL coding: *I01: <<elle s'appelle>[/] elle s'appelle>[@fra] <ouh what am I doing?>[@eng]. [+ fe]

It is important to point out that while these changes would now enable FREQ, VOCD and WDLEN (and potentially many other CLAN commands) to be used accurately, they would not compromise the original needs of the research teams – to be able to automatically exclude the data they were not interested in, i.e the English material.

A further suggestion for corpora already transcribed in CHAT format would be to insert an addressee tier for each utterance. The advantages of doing this have been constantly emphasised throughout this study. However, such insertion could be problematic, or even unfeasible, in transcripts of interactions in which there are more than two interlocutors. Without detailed contextual information it may be impossible to accurately determine the addresse(s) of every single utterance and even access to original audio recordings would be of little help (video recordings would be of greater help here). Of course, for those researchers who are in the process of compiling new corpora, a recommendation would be to require those collecting the spoken data to make detailed addressee notes so that addressee tiers could be accurately inserted when transcribing the data.

Although the discussion above has focussed on corpora already transcribed in CHAT format, I would like to argue that the time expended on converting existing bilingual corpora into CHAT format and on inserting the language coding recommended in my study would be worthwhile, offering substantial rewards in terms of new insights. For example, a set of corpora which I believe would benefit from my methodological recommendations is that collected by John Lipski, currently Professor of Spanish and Linguistics at Pennsylvania State University in the U.S.A and whose main research interests lie in language contact phenomena[209]. To illustrate the benefits of my specific corpus linguistics approach to investigating code-switching, I will make reference to Lipski's recent study (2014) in which he carried out a componential analysis of code-switching in two of his corpora, the Texas corpus (containing data from fluent Mexican-American bilinguals) and the NW Louisiana corpus (consisting of data from low fluency American bilingual heritage speakers of Spanish).

[209] See www.personal.psu.edu/jm134/ for more details of Lipski's research interests.

In the above-mentioned study he extracted code-switches from both corpora (324 from the former and 160 from the latter). Although he does not detail his method of extraction I assume this was done mannually, especially as he mentions that he only used a small portion of his data (the first two hours of 30 hours of the spoken material in the Texas corpus[210]). Using Muysken's typology (2000) to categorize the switches, Lipski found quantitative and qualitative differences in the code-switches used by the two sets of speakers: for the fluent bilinguals of the Texas corpus alternation predominated (80% of the switches were of this type) whereas the Louisiana heritage speakers' code-switches were characterised by congruent lexicalization (60%). He proposed that his findings would support an amendment to Muysken's typology with regards to the latter code-switching type: 'congruent lexicalization' could be characterised as being 'fluent' or 'low fluency and possibly involuntary' (see Table 7 in Lipski, 2014:21 for further details). He concludes that future research should look at investigating low-fluency code-switching in second-language learners and non-balanced heritage language bilinguals, in order to examine the nature of each language's contribution to the bilingual utterances produced by such speakers and thus determine 'the precise relationship between non-constituent congruent lexicalization and language dominance' (2014:44).

I would like to suggest that if Lipski were able to replicate the types of quantitative analyses that I have reported on in this dissertation, his data would provide him with the insights he is seeking. As demonstrated through the quantitative examination of the LOBILL Corpus, FREQ, VOCD and WDLEN provide the means to empirically measure the differential contribution of each language to code-switched utterances and determine whether, and how, one language plays a more dominant role than the other. In addition, the resulting frequency word lists would provide the means for Lipski to examine the grammatical and lexical nature of each language's contribution to the code-switches occurring in his data. Furthermore, in contrast to the study discussed above in which Lipski restricted his selection of data (presumably for practical reasons), he would be able to perform analyses on all of the data in any of his corpora at the touch of a button. There are clear advantages of being able to make use of more data: the potential for spotting patterns in the output

[210] As he only refers to the Texas Corpus, one assumes that he made use of all of the data in the NW Louisiana corpus.

is greater and more output would mean more evidence to support any conclusions or hypotheses.

I would hope that Lipski, and others like him who are in possession of valuable bilingual corpora, would be encouraged to contribute their data to the wider research community in a format which would facilitate and maximise their exploitability (such as that used in the LOBILL Corpus). Although it is evident that there may be good reasons why some existing corpora cannot be made available for others to analyse (ethical restrictions, for example), I feel that at least greater transparency in terms of the reporting of methodology would be of benefit to the research community. For example, although Poplack calls for more quantitative studies of code-switching (see earlier reference in Chapter 1), few, if any, details of the methodology she uses in her own quantitative studies (carried out on large French corpora which are not freely available[211]), are given. As a result potential methodological insights are lost, both for the reader and Poplack herself: the reader is not provided with the means to replicate her quantitative approach and this means that Poplack cannot benefit from any suggestions others might have regarding her methodology. By ensuring methodological transparency in my study I hope to offer the code-switching research community insights into how fruitful a corpus-based approach such as mine can be. I also hope to benefit from the comments of other researchers who may be able to provide me with suggestions for further improvements to my methods of transcription and analysis and/or alterations to the schema I propose for the interpretation of quantitative measures when used to characterise the roles of languages participating in bilingual utterances. As for the LOBILL Corpus itself, I have briefly hinted at the linguistic phenomena which could be investigated in the data. However, I would welcome further suggestions with regards to its exploitability and look forward to the possibility of collaborative studies based on this unique longitudinal corpus of spoken child bilingual language.

[211] Details of these corpora can be found at www.sociolinguistic.uottawa.ca/holdings.html and a comprehensive list of publications are listed under the following link: www.sociolinguistic.uottawa.ca/publications.html

References

Adolphs, S., & Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Routledge.

Allen, S., Genesee, F., Fish, S. & Crago, M. (2001). Patterns of Code Mixing in English-Inuktitut Bilinguals. In *Papers From the Annual Meeting of the Chicago Linguistic Society.* CLS 37.2: The Panels, 171-188.

Ambridge, B. (2013). How Do Children Restrict Their Linguistic Generalizations? An (Un-) Grammaticality Judgment Study. *Cognitive science*, *37*(3), 508-543.

Auer, P. (1995). The pragmatics of code-switching: a sequential approach. In L. Milroy & P. Muysken (Eds.), *One Speaker, Two Languages: cross-disciplinary perspectives on codeswitching* (pp.115-135). Cambridge: Cambridge University Press.

Backus, A. & Dorleijn, M. (2009). Loan translations versus code-switching. In B. E. Bullock & J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp.170-187). Cambridge: Cambridge University Press.

Barron-Hauwaert, S. (2004). *Language strategies for bilingual families: The one-parent-one-language approach*. Clevedon: Multilingual Matters.

Belazi, H. M., Rubin, E. J. & Toribio, A. J. (1994). Code Switching and X-Bar Theory: The Functional Head Constraint. *Linguistic Inquiry*, 25 (2).

Bentz, C., & Buttery, P. (2014, April). Towards a computational model of grammaticalization and lexical diversity. In *Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)@ EACL* (pp. 38-42).

Berko Gleason, J. (1975). Fathers and other strangers: Men's speech to young children. In *Georgetown University Roundtable on Language and Linguistics* (pp.289-297). Washington D.C.: Georgetown University Press.

Bernardini, P. & Schlyter, S. (2004). Growing syntactic structure and code-mixing in the weaker language: The ivy hypothesis. *Bilingualism: Language and Cognition,* 7(1), 49-69.

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., ... & Zesch, T. (2013). Scalable construction of high-quality web corpora. *Special issue of JLCL. To appear*.

Blom, J-P. & Gumperz, J. (1972). Social Meaning in Linguistic Structures: Code Switching in Northern Norway. In J. Gumperz & Del Hymes (Eds.). *Directions in Sociolinguistics: The Ethnography of Communication* (pp.407-434). New York: Holt, Rinehart, and Winston.

Bloomfield, L. (1933, reprinted 1984). *Language*. New York: Holt, Rinehart & Winston Inc.

Blum-Kulka, S. (2008). Language socialization and family dinnertime discourse. In P. Duff, & N. Hornberger (Eds.). *Encyclopedia of Language and Education, Vol. 8 Language Socialization* (pp.87-99). New York: Springer.

Bolonyai, A. (2009). Code-switching, imperfect acquisition, and attrition. In B. E. Bullock & J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp.255-287). Cambridge: Cambridge University Press.

Brosseau-Lapré, F., & Rvachew, S. (2013). Cross-linguistic comparison of speech errors produced by English-and French-speaking preschool-age children with developmental phonological disorders. *International journal of speech-language pathology*, (0), 1-11.

Brulard, I., & Carr, P. (2003). French-English bilingual acquisition of phonology: One production system or two?. *International Journal of Bilingualism*, *7*(2), 177-202.

Bullock, B. E. & Toribio, J. (Eds.), (2009). *The Cambridge Handbook of Linguistic Code-switching*. Cambridge: Cambridge University Press.

Cantone, K., F. (2007). *Code-switching in Bilingual children: studies in theoretical psycholinguistics*, v.37. Dordrecht: Springer.

Cantone, K. F. & MacSwan, J. (2009). Adjectives and word order: A focus on Italian-German code-switching. In L. Isurin et al. (Eds.), *Studies in Bilingualism Volume 41: Multidisciplinary Approaches to Code Switching*. (pp.243-277). Amsterdam/Philadelphia: John Benjamins Company:

Carr, P. (2007). Internalism, externalism and coding. *Language Sciences*, *29*(5), 672-689.

Carter, D., Deuchar, M., Davies, P., & Couto, M. D. C. P. (2011). A systematic comparison of factors affecting the choice of matrix language in three bilingual communities. *Journal of Language Contact*, *4*(2), 153-183.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs and P. Rosenbaum (Eds.) *Readings in English transformational grammar*,). Waltham, MA:Ginn.

Chomsky, N. (1981). *Lectures on government and binding*. New York: Mouton de Gruyter.

Chomsky, N. (1991). Linguistics and cognitive science: Problems and mysteries. In A. Kasher (Ed.), *The Chomskyan turn.* Cambridge: Blackwell.

Chomsky, N. (1994). Bare phrase structure. *MIT occasional papers in linguistics* 5. Also published in G. Webelhuth (Ed.) (1995) *Government and Binding Theory and the Minimalist Program*. Oxford: Blackwell.

Clyne, M. (1967). *Transference and triggering*. The Hague: Nijhoff.

Clyne, M. (1987). Constraints on CS: how universal are they? *Linguistics* 25, 739-764.

Coyle, Y., & Roca de Larios, J. (2013). Exploring the role played by error correction and models on children's reported noticing and output production in a L2 writing task. *Studies in Second Language Acquisition*, 1-35.

Cruz-Ferreira, M. (1999). Prosodic mixes: Strategies in multilingual language acquisition. *International Journal of Bilingualism*, *3*(1), 1-21.

Cruz-Ferreira, M. (2006). *Three is a crowd?: Acquiring Portuguese in a trilingual environment* (Vol. 6). Multilingual Matters.

Cruz-Ferreira, M. (2010). *Multilinguals Are--?*. London: Battlebridge Publications.

De Houwer, A. (1990). The Acquisition of Two Languages from Birth: a case study. Cambridge: CUP.

De Houwer, A. (2009). Bilingual First Language Acquisition. Clevedon: Multilingual Matters.

Deuchar, M., & Clark, A. (1996). Early bilingual acquisition of the voicing contrast in English and Spanish. *Journal of Phonetics*, *24*(3), 351-365.

Deuchar, M. & Quay, S. (2000). Bilingual Acquisition: theoretical implications of a case-study. Oxford: OUP.

Dewaele, J. M. (2001). Activation or inhibition? The interaction of L1, L2 and L3 on the language mode continuum. *Bilingual Education and Bilingualism*, 69-89.

Dewaele, J. M. (2007). Still trilingual at ten: Livia's multilingual journey.*Multilingual Living Magazine*, 68-71.

Di Scuillo, A-M., Muysken, P. & Singh, R. (1986). Government and Code-Switching. *Journal of Linguistics*, 22, 1-24.

Dodd, B. (2013). *Differential diagnosis and treatment of children with speech disorder*. John Wiley & Sons.

Duff, P. & Hornberger, N. (Eds.) (2008). *Encyclopedia of Language and Education, Vol. 8 Language Socialization.* New York: Springer.

Duran, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220-242.

Edwards, M. & Gardner-Chloros, P. (2007). Compound verbs in Code-Switching: bilinguals making do? *International Journal of Bilingualism*. 11(1), 73-91.

Eichler, N., Jansen, V., & Müller, N. (2013). Gender acquisition in bilingual children: French–German, Italian–German, Spanish–German and Italian–French. *International Journal of Bilingualism*, *17*(5), 550-572.

Endesfelder-Quick, A. (2013) Mixing within the NP – comparing German-English and German-Russian bilingual children. Paper presented at the conference 'Code-switching in the bilingual child: within and across the clause' held at Bergische Universitat, Wuppertal, Germany between 18th and 20th April 2013.

Gagarina, N. V. (2013). Acquisition and loss of L1 in a Russian-German bilingual child: A case study. *Путь в язык. Одноязычие и двуязычие*, 137.

Gardner-Chloros, P. (1997). Code-Switching: Language Selection in three Strasbourg department stores. In N. Coupland & A. Jaworski. (Eds.), Sociolinguistics: a reader and coursebook. (pp.361-376). Basingstoke/London: Macmillan.

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

Gardner-Chloros, P. Moyer, M & Sebba, M. (2000). The LIDES Coding Manual: A document for preparing and analyzing language interaction data. Prepared jointly with the LIPPS Group (+11 contributors). *International Journal of Bilingualism*. 4(2), 131-270.

Gathercole, V. C. M., Thomas, E. M. & Hughes, E. K (2008) Designing a normed receptive vocabulary test for bilingual populations: A model from Welsh. *International Journal of Bilingual Education and Bilingualism*. 11(6), 678-720.

Gathercole, V. C. M., Thomas, E. M., Roberts, E., Hughes, C., & Hughes, E. K. (2013). Why assessment needs to take exposure into account: Vocabulary and grammatical abilities in bilingual children. *Issues in the Assessment of Bilinguals*, 20-55.

Genesee, F. (2006). Bilingual First Language Acquisition in Persepective. In P. McCardle & E. Hoffs (Eds.), *Childhood Bilingualism: Research on infancy through school age* (pp.47-61). Clevedon: Multilingual Matters.

Gildersleeve-Neumann, C., & Goldstein, B. A. (2014). Cross-linguistic generalization in the treatment of two sequential Spanish-English bilingual children with speech sound disorders. *International journal of speech-language pathology*, (0), 1-15.

Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Perez, A. (2013). Identification of Specific Language Impairment in Bilingual Children: I. Assessment in English. *Journal of Speech, Language, and Hearing Research*,*56*(6), 1813-1823.

Gollan, T. H., Schotter, E. R., Gomez, J., Murillo, M., & Rayner, K. (2014). Multiple Levels of Bilingual Language Control Evidence From Language Intrusions in Reading Aloud. *Psychological science*, *25*(2), 585-595.

Granger, S. (2013). Error-tagged learner corpora and CALL: A promising synergy. *CALICO journal*, *20*(3), 465-480.

Gullberg, M., Indefrey, P. & Muysken, P. (2009). Research techniques for the study of code-switching. In B. E. Bullock & J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp.21-39). Cambridge: Cambridge University Press.

Hager, M., Schmeißer, A. and Bucher, M. (2013). Influencing factors on code-switching in a cross-sectional study with German-Romance (Spanish, French, Italian) bilingual children. Paper presented at the conference 'Code-switching in the bilingual child: within and across the clause' held at Bergische Universitat, Wuppertal, Germany between 18th and 20th April 2013.

Hernandez, A. E. (2013). Neural and psychological bases of switching in bilingual children. Paper presented at the conference 'Code-switching in the bilingual child: within and across the clause' held at Bergische Universitat, Wuppertal, Germany between 18th and 20th April 2013.

Hickey, R. (Ed.), (2010). *The Handbook of Language Contact*. Oxford:Wiley-Blackwell.

Hoff, E., Core, C., Place, S., Rumiche, R., Senor, M. and Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*. 39(1), 1-27.

Isurin, L., Winford, D., & De Bot, K. (Eds.). (2009). *Multidisciplinary approaches to code switching* (Vol. 41). John Benjamins Publishing.

Jaeger, J. J. (2013). *Kids' slips: What young children's slips of the tongue reveal about language development*. Psychology Press.

Jake, J. L. & Myers-Scotton, C. (2009). Which language? Participation potentials across lexical categories in code-switching. In L. Isurin et al. (Eds.), *Studies in Bilingualism Volume 41: Multidisciplinary Approaches to Code Switching*. (pp.207-242). Amsterdam/Philadelphia: John Benjamins Company:

Jake, J., Myers-Scotton, C. & Gross, S. 2002. Making a minimalist approach to codeswitching work: adding the Matrix Language. *Bilingualism: Language and Cognition*, 5(1): 69-91.

James, C. (2013). *Errors in language learning and use: Exploring error analysis*. Routledge.

Jarvis, S. (2009). Lexical Transfer. In A. Pavlenko (Ed.), *The Bilingual Mental Lexicon: interdisciplinary Approaches* (pp.99-124). Bristol: Multilingual Matters.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Joshi, A. (1985). Processing of Sentences with Intrasential Code Switching. In D. R. Dowty, L. Kattunen & A. M. Zwickey. (Eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. Cambridge: Cambridge University Press.

Juan-Garau, M., & Perez-Vidal, C. (2001). Mixing and pragmatic parental strategies in early bilingual acquisition. *Journal of child language*, *28*(01), 59-86.

Koppe, R. (in press). Is code switching acquired? In J. MacSwan (Ed.) *Grammatical Theory and Bilingual Code Switching.* Cambridge, MA: MIT Press.

Lanza, E. (1997). *Language Mixing in Infant Bilingualism: a sociolinguistic perspective*. Oxford: Oxford University Press.

Lanza, E. (2007). *Language Mixing in Infant Bilingualism: a sociolinguistic perspective*. Oxford: Oxford University Press.

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. *Applied linguistics*, *12*(2), 180-196.

Leopold, W. (1939-1949). *Speech Development of a Bilingual Child.* Evanston, Illinois: Northwestern University Press, 4 volumes.

Leopold, W. (1970, 1939-1949). *Speech Development of a Bilingual Child. A linguist's record.* New York: AMS Press.

Li Wei. (2002). 'What do you want me to say?' On the Conversation Analysis approach to bilingual interaction. *Language in Society*. 32, 159-180.

Li Wei. (2005). "How can you tell?" Towards a common sense explanation of conversational code-switching. *Journal of Pragmatics*. 37(3), 375-389.

Liceras, J. (2013) Gender agreement patterns in mixed concord and agreement structures: does 'code-switching' matter? Paper presented at the conference 'Code-switching in the bilingual child: within and across the clause' held at Bergische Universitat, Wuppertal, Germany between 18th and 20th April 2013.

Lipski, J. (1985). *Linguistic Aspects of Spanish-English Language-Switching*. Center for Latin American Studies, Arizona State University.

Lipski, J. M. (2014). Spanish-English code-switching among low-fluency bilinguals: Towards an expanded typology. *Sociolinguistic Studies*, *8*(1), 23-55.

Liu, Y. (2008). Evaluation of the Matrix Language Hypothesis: Evidence from Chinese-English Code-switching Phenomena in Blogs. *Journal of Chinese Language and Computing*, *18*(2), 75-92.

MacSwan, J. (1997). *A minimalist approach to intrasentential Code Switching: Spanish-Nahuatl Bilingualsim in Central Mexico*. (PhD dissertation) http://www.public.asu.edu/~macswan/macswan.pdf (accessed 29.11.2010)

MacSwan, J. (1999). *A Minimalist Approach to Intrasentential Code Switching*. New York, Garland.

MacSwan, J. (2000). The architecture of the bilingual language faculty: Evidence from codeswitching. *Bilingualism: Language and Cognition*. 3(1), 37-54.

MacSwan, J. (2005a). Codeswitching and generative grammar: A critique of the MFL model and some remarks on "modified minimalism". *Bilingualism: Language and Cognition* 8(1), 1-22.

MacSwan, J. (2005b). Precis of a Minimalist Approach to Intrasentential Code Switching. *Italian Journal of Linguistics* 17(1), 55-92.

MacSwan, J. (2013). Paper presented at the conference 'Code-switching in the bilingual child: within and across the clause' held at Bergische Universitat, Wuppertal, Germany between 18th and 20th April 2013.

MacWhinney, B. (1991). *The CHILDES Project: tools for Analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2014a). The Database Manuals. http://childes.psy.cmu.edu/manuals .

MacWhinney, B. (2014b). *The CHILDES Project, Tools for Analyzing Talk – Electronic Edition. Part 1: The CHAT Transcription Format*. Carnegie Mellon University. Available online: http://childes.psy.cmu.edu/manuals/chat/pdf .

MacWhinney, B. (2014c). *The CHILDES Project, Tools for Analyzing Talk – Electronic Edition. Part 2: The CLAN Programs.* Carnegie Mellon University. Available online: http://childes.psy.cmu.edu/manuals/clan/pdf .

MacWhinney, B. & Snow, C. (1985) The child language data exchange system. *Journal of Child Language*, 12, 271-296.

MacWhinney, B. & Snow, C. (1990) The child language data exchange system: an update. *Journal of Child Language*, 17, 457-472.

Mahootian, S. (1993). *A Null Theory of Codeswitching*. Ph.D. dissertation. Northwestern University.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, Basingstoke: Palgrave Macmillan.

Matras, Y. (2010). Contact, Convergence, and Typology. In R. Hickey (Ed.), *The Handbook of Language Contact* (pp.66-85). Oxford:Wiley-Blackwell.

McCarthy, P. H. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488.

Meisel, J. (1994). Codeswitching in young bilingual children. The acquisition of grammatical constraints. *Studies in Second Language Acquisition* 16, 413-439.

Morris, D. & Jones, K. (2008). Language socialization in the home and minority language revitalization in Europe. In P. Duff, & N. Hornberger (Eds.). *Encyclopedia of Language and Education, Vol. 8 Language Socialization* (pp.127-143). New York: Springer.

Muller, N. & Cantone, K. F. (2009). Language mixing in bilingual children: code-switching? In B. Bullock & J. Toribio. (Eds). *The Cambridge Handbook of Linguistic Code-switching* (pp.200-219). Cambridge: Cambridge University Press.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Myers-Scotton, C. M. (1993a). *Duelling languages: Grammatical structure in codeswitching.* Oxford: Oxford University Press.

Myers-Scotton, C. M. (1993b). *Social motivations for codeswitching. Evidence from Africa.* Oxford: Oxford University Press.

Myers-Scotton, C. M. (2002). *Contact Linguistics: bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.

Myers-Scotton, C. & Jake, J. L. (2009). A universal model of code-switching and bilingual language processing and production. In B. Bullock & J. Toribio. (Eds). *The Cambridge Handbook of Linguistic Code-switching*. Cambridge: Cambridge University Press.

Ochs, E. (1985). Variation and error: A sociolinguistic study of language acquisition in Samoa. In D. Slobin (Ed.). *The Cross-Linguistic Study of Language Acquisition*, Vol. 1, 783-838. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ochs, E. & Schieffelin, B. (1984). Language acquisition and socialization: Three developmental stories. In R.A. Shweder and R.A. LeVine (Eds.), *Culture Theory: Essays on Mind, Self, and Emotion* (pp.276-320). Cambridge: Cambridge University Press.

Ochs, E. & Schieffelin, B. (1995). The impact of language socialization on grammatical development. In P. Fletcher & B. MacWhinney (Eds.). *The Handbook of Child Language* (pp.73-94). *Oxford:* Blackwell.

Ochs, E. & Schieffelin, B. (2008). Language Socialization: An Historical Overview. In P. Duff, & N. Hornberger (Eds.). *Encyclopedia of Language and Education, Vol. 8 Language Socialization* (pp.4-15). New York: Springer.

Ochs, E. & Sholet, M. (2006). The cultural structuring of mealtime socialization. In R. Larson, A. Wiley, & K. Branscomb (Eds.). *Family mealtime as a context of development and socialization. New Directions in Child and Adolescent Development Series*, Vol. 11, 35-50. Jossey-Bass, San Franscisco.

Ochs, E. & Taylor, C. (1992). Family narrative as political activity. *Discourse & Society* 3, 301-341.

Paradis, J., Nicoladis, E. & Genesee, F. (2000). Early emergence of structural constraints on code-mixing: Evidence from French-English bilingual children. *Bilingualism: Language and Cognition*, 3(3), 245-261.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*,*108*(9), 3526-3529.

Pittman, I. (2008). Bilingual and trilingual codeswitching between Hungarian, Romanian and English in the speech of two Transylvanians living in North America. *International Journal of Multilingualism*, *5*(2), 122-139.

Plaff, C. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 55:291-318.

Poplack, S. (1980). 'Sometimes I'll Start a Sentence in Spanish Y TERMINO EN ESPAÑOL': Toward a Typology of Code-Switching. *Linguistics,* 18, 581-618.

Poplack, S. (1981). Syntactic structure and social function of code-switching. In R. P. Duran (Ed.), *Latino Discourse and Communicative Behaviour* (pp.169-184). New Jersery: Ablex Publishing Corporation.

Poplack, S. (2001). Code-switching (linguistic). *International encyclopedia of the social and behavioral sciences*, 2062-2065.

Poplack, S. (2013). Toward a typology of code-switching. *Linguistics*, 51 (Jubilee), 11-14. De Gruyter Mouton.

Poplack, S., & Turpin, D. (2011). O Futur tem futuro no frances (Canadense)? *Cadernos de Estudos Lingüísticos*, *36*.

Poplack, S., & Dion, N. (2012). Myths and facts about loanword development. *Language Variation and Change*, *24*(03), 279-315.

Prinz, P. M., Pemberton, E., & Nelson, K. E. (1985). The ALPHA interactive microcomputer system for teaching reading, writing, and communication skills to hearing-impaired children. *American Annals of the Deaf*, *130*(5), 444-461.

Prinz, P. M., & Prinz, E. A. (1979). Simultaneous acquisition of ASL and spoken English (in a hearing child of a deaf mother and hearing father): Phase I: Early lexical development. *Sign Language Studies*, *25*(1), 283-296.

Räsänen, S. H., Ambridge, B., & Pine, J. M. (2013). Infinitives or bare stems? Are English-speaking children defaulting to the highest-frequency form?. *Journal of child language*, 1-24.

Richards, B. J., & Malvern, D. D. (2007). Validity and threats to the validity of vocabulary assessment. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 79–92). Cambridge: Cambridge University Press.

Ronjat, J. (1913). *Le Développement du langage observé chez un enfant bilingue*. Paris: Champion.

Sankoff, D. & Poplack, S. (1981). A Formal Grammar of Code-Switching. *Papers in Linguistics*, 14, 3-45.

Sauve, D. & Genesee, F. (2000, March). *Grammatical constraints on child bilingual code-mixing.* Paper presented at the Annual Conference of the American Association for Applied Linguistics, Vancouver, Canada.

Schieffelin, B.B. (1985). The acquisition of Kuali. In D. Slobin (ed.), *The Cross-Linguistic Study of Language Acquisition*, Vol. 1, 252-593. New Jersey: Lawrence Erlbaum Associates, Hillsdale.

Schmid, M. S. (2010). Languages at play: The relevance of L1 attrition to the study of bilingualism. *Bilingualism: Language and Cognition,* 13 (1), 1-7.

Schmid, M. S. & Kopke, B. (2009). Attrition and the Mental Lexicon. In A. Pavlenko (Ed.), *The Bilingual Mental Lexicon: Interdisciplinary Approaches*. (pp.210-228). Bristol: Multilingual Matters.

Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency–Zipf revisited. *Studia Linguistica*, *58*(1), 37-52.

Smith, R. (2012). Distinct word length frequencies: distributions and symbol entropies. *Glottometrics*, 23, 7-22.

Snow, C. (1991). The theoretical basis for relationships between language and literacy development. *Journal of Research in Childhood education* 6 (Fall/Winter), 5-10.

Stocco, A., Yamasaki, B., Natalenko, R., & Prat, C. S. (2014). Bilingual brain training: A neurobiological framework of how bilingual experience improves executive function. International Journal of Bilingualism. Vol. 18 (1), 67-92

Stroud, C. (1992). The problem of definition and meaning in code-switching. *Text*, 12(1), 127-155.

Stroud, C. (1998). Perspectives on cultural variability of discourse and some implications for codeswitching. In P. Auer (Ed.) *Codeswitching in Conversation: Language, Interaction and Identity* (pp.321-348*)*. London: Routledge.

Thomason, S.G. & Kaufman, T. (1988). *Language Contact, Creolization and Genetic Linguistics.* Berkeley: University of California Press.

Torruella, J & Capsada, R. (2013). Lexical statistics and typological structures: measure of lexical richness. Paper given at V International Conference on Corpus Linguistics (CILC2013). Procedia. Available on-line at www. sciencedirect.com

Travis, C. E. & Torres Cacoullos, R. (2013) Making voices count: Corpus compilation in bilingual communities. A*ustralian Journal of Linguistics*

Treffers-Daller, J. (2009a). Language dominance and lexical diversity: How bilinguals and L2 learners differ in their knowledge and use of French lexical and functional items. In B. Richards, H. M. Daller, D. D. Malvern, P. Meara, J. Milton and J. Treffers-Daller (Eds.) *Vocabulary studies in first and second lan guage acquisition. The interface between theory and applications* (pp. 74–90). Houndmills Basingstoke: Palgrave Macmillan.

Treffers-Daller, J. (2009b). Code-switching and transfer: an exploration of similarities and differences. In B. E. Bullock & J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp.58-74). Cambridge: Cambridge University Press.

Treffers-Daller, J. (2010). Operationalizing and measuring language dominance. *International Journal of Bilingualism.* 15(2) 147–163 Sage Publications.

Treffers-Daller, J., & Korybski, T. (2015). Using lexical diversity measures to operationalise language dominance in bilinguals. In: Silva-Corvalan, C. and Treffers-Daller, J. (eds). Language dominance in bilinguals: issues of measurement and operationalization. Cambridge University Press, Cambridge. (In Press). Available at http://centaur.reading.ac.uk/39019/.

Vihman, M. (1985). Language differentiation by the bilingual infant. *Journal of Child Language*. 12, 297-324.

Vihman, M. (1998). A developmental perspective on codeswitching: Conversations between a pair of bilingual siblings. *International Journal of Bilingualism*, *2*, 45-84.

Winford, D. (2010). Contact and Borrowing. In R. Hickey (Ed.), *The Handbook of Language Contact* (pp.170-187). Oxford:Wiley-Blackwell.

Woolard, K. A. (1997). Between friends: gender, peer group structure and bilingualism in urban Catalonia. *Language in Society*. 26, 533-560.

Woolford, E. (1983). Bilingual code-switching and syntactic theory. *Linguistic Inquiry*, 14(5), 520-36.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., ... & Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, 0142723711409976.

Yip, V., & Matthews, S. (2000). Syntactic transfer in a Cantonese–English bilingual child. *Bilingualism: Language and cognition*, *3*(03), 193-208.

Yip, V. & Matthews, S. (2007). *The Bilingual Child: Early Development and Language Contact*. Cambridge: Cambridge University Press.

Zentella, A.C. (1997). *Growing up Bilingual: Puerto Rican Children in New York*. Oxford/Malden, MA: Blackwell.

Zipf, G. K. (1935). The psycho-biology of language.

Appendix A. List of files in the LOBILL Corpus

File name	Interlocutors (and age of siblings)	Location and activity
001CHenJ&MAUG01	JAM (3;5.18) MEG (5;10.12) MOT, GRA, BEC, PAI	At home, Fortaleza, Brazil. Opening presents in the living room.
002CHenJ&MAUG01	JAM (3;5.24) MEG (5;10,18) MOT, GRA, BEC	At friends' house, Pacoti, Brazil.Talking and eating hotdogs.
003INenMSEP01	JAM (3;6.6) MEG (5;11.0), MOT	At home, Fortaleza, Brazil. Mother asking Meggie questions.
004FPenJ&MSEP01	JAM (3;6.10) MEG (5;11.4), MOT	At home, Fortaleza, Brazil. Playing with lego bricks.
005INenMSEP01	MEG (5;11.18), MOT	At home, Fortaleza, Brazil. Mother asking Meggie about words.
006LAenMOCT01	JAM (3;7.18) MEG (6;1.12), MOT	At home, Fortaleza, Brazil. Meggie talking about the the video of 'Lady and the Tramp'.
007FPenMNOV01	MEG (6;2.1), SAR, MOT	At friends' house, Pacoti, Brazil. Mixing colours for painting.
008LAen MJUN02	MEG (6;8.15), MOT	In the car on the way to the beach house in Redonda, Brazil. Meggie reading and talking about dinosaurs.
009CHenJJUN02	JAM (4;3.21), MOT	At the beach house in Redonda, Brazil. James and mother chatting about the day.
010CHenJ&MJUN02	JAM (4;3.21) MEG (6;8.15), MOT	At the beach house in Redonda, Brazil. Chatting about the São João festival.
011FPenJ&MJUN02	JAM (4;3.27) MEG (6;8.21), MOT	At home, Fortaleza, Brazil. Playing in the bedroom with dolls.
012INenMJUN02	MEG (6;8.22), MOT	At home, Fortaleza, Brazil. Mother asking Meggie questions on the day Brazil won the World Cup.
013MTenJ&MJUL02	JAM (4;4.1) MEG (6;8.25), MOT	At home, Fortaleza, Brazil. Having lunch.
014MTenJ&MJUL02	JAM (4;4.2) MEG (6;8.28), MOT	At home, Fortaleza, Brazil. Having lunch.
015MTenJ&MJUL02	JAM (4;4.6) MEG (6;9.1), MOT, PAI	At home, Fortaleza, Brazil. Having lunch.
016INenJUL02	JAM (4;4.26), MOT	In the car on the way to the beach house in Redonda, Brazil. Mother asking James questions.
017MTenJ&MAUG02	JAM (4;5.1) MEG (6;9.25), MOT, PAI	At home, Fortaleza, Brazil. Having lunch.
018FPenJAUG02	JAM (4;5.2), MOT	At home, Fortaleza, Brazil.James counting coins in the living room.
019LAenJ&MAUG02	JAM (4;5.6)	At the beach house in Redonda, Brazil. Meggie

	MEG (6;10.0), MOT	recounting the story of Postman Pat (from the pictures).
020MTenJ&MAUG02	JAM (4;5.8) MEG (6;10.2), MOT	At home, Fortaleza, Brazil. Having lunch.
021MTenJ&MAUG02	JAM (4;5.14) MEG (6;10.8), MOT	At home, Fortaleza, Brazil. Having lunch.
022PGenJ&MAUG02	JAM (4;5.14) MEG (6;10.8), MOT	At home, Fortaleza, Brazil. Playing board game 'Days of the week' in the living room.
023MTenJ&MAUG02	JAM (4;5.24) MEG (6;10.18), MOT	At home, Fortaleza, Brazil. Having lunch.
024LAenJ&MOCT02	JAM (4;7.24) MEG (7;0.18), MOT	At home, Fortaleza, Brazil. Meggie is recounting a dream.
025FPenJ&MOCT02	JAM (4;7.18) MEG (7;0.18), MOT	At home, Fortaleza, Brazil. James and mother making a house out of lego bricks.
026INenMNOV02	MEG (7;1.6), MOT	At home, Fortaleza, Brazil. Mother asking Meggie questions.
027PGenJ&MNOV02	JAM (4;8.19) MEG (7;1.13), MOT	At home, Fortaleza, Brazil. Playing 'Donkey' card game in the bedroom.
028PGenMDEC02	MEG (7;2.8), MOT, PAI	At home, Fortaleza, Brazil. Meggie and mother are playing a guessing game with animal biscuits.
029PGenJDEC02	JAM (4;9.15), MOT	At home, Fortaleza, Brazil. Playing guessing game with animal cards.
030CHenJ&MDEC02	JAM (4;9.24) MEG (7;2.18), MOT, GRA, PAI	At home, Fortaleza, Brazil. Opening Christmas presents.
031CHenJDEC02	JAM (4;9.30), MOT, GRA	At home, Fortaleza, Brazil. James recounting to Grandma the incident of the punctured tyre.
032CHenJ&MJAN03	JAM (4;10.0) MEG (7;2.23), MOT, PAI, GRA	At the airport on the plane viewing area, Fortaleza, Brazil. i) watching the planes with Grandma ii) watching the planes after having said goodbye to Grandma iii) watching Grandma's plane take off iv) watching other planes take off
033FPenMFEB03	MEG (7;4.9), MOT	At home, Fortaleza, Brazil. Meggie playing in the bedroom.
034LAenJ&MFEB03	JAM (4;11.16) MEG (7;4.9), MOT	At home, Fortaleza, Brazil. James doing homework in the study.
035MTenJ&MFEB03	JAM (4;11.21) MEG (7;4.15), MOT	At home, Fortaleza, Brazil. James having lunch while Meggie makes jewellery.
036FPptJ&MMAR03	JAM (5;0.1) MEG (7;4.25), MOT, SAR, VOV, PAI	At the beach house in Redonda, Brazil. James, Meggie and Sara are making plaster cast moulds of Mickey mouse, helped by adults.
037FPenJMAR03	JAM (5;0.1), MOT	At the beach house in Redonda, Brazil. James is playing outside with the sand.
038PGenJ&MMAR03	JAM (5;0.3)	At the beach house in Redonda, Brazil. In the

	MEG (7;4.27), MOT, PAI	bedroom, listening to the storm and playing 'Uno' card game.
039MTenJ&MMAR03	JAM (5;0.13) MEG (7;5.7), MOT, PAI	At home, Fortaleza, Brazil. Having lunch.
040CHenJMAR03	JAM (5;0.18), MOT	At home, Fortaleza, Brazil. In kitchen talking about the overnight school trip.
041LAenMMAR03	MEG (7;5.19), MOT	At home, Fortaleza, Brazil. Meggie reading 'The ugly duckling' out loud.
042CHenJ&MAPR03	JAM (5;1.7) MEG (7;6.1), MOT	At home, Fortaleza, Brazil. James talking to mother about school activities.
043CHenJAPR03	JAM (5;1.16), MOT	At the beach house in Redonda, Brazil. In the bedroom at night James talking to mother.
044PGMAPR03	MEG (7;6.11), MOT, PAI	At the beach house in Redonda, Brazil. Playing 'Uno' card game.
045LAenJ&MAPR03	JAM (5;1.18) MEG (7;6.12), MOT	At the beach house in Redonda, Brazil. Lying in a hammock, Meggie reading out loud a story that she wrote about her guinea pig Biju.
046FPenJMAY03	JAM (5;2.0), MOT	At home, Fortaleza, Brazil. James making a car out of lego bricks in the living room.
047PGptJ&MMAY03	JAM (5;2.0) MEG (7;6.24), SAR, MOT	At home, Fortaleza, Brazil. Playing 'Donkey' card game in the bedroom.
048INenJMAY03	JAM (5;2.16), MOT	At home, Fortaleza, Brazil. Mother asking James questions about forthcoming trip to England.
049INenMMAY03	MEG (7;7.16), MOT	At home, Fortaleza, Brazil. Mother asking Meggie questions about forthcoming trip to England
050MTenJ&MMAY03	JAM (5;2.22) MEG (7;7.16), MOT, PAI	At home, Fortaleza, Brazil. Having pizza in the kitchen.
051PGenMJUN03	MEG (7;7.28), MOT	At home, Fortaleza, Brazil. Playing dominoes before going to bed.
052CHptJ&MJUN03	JAM (5;3.10) MEG (7;8.4), MOT, PAI, SAR, JUL	At the airport, Fortaleza, Brazil. Chatting to cousins before getting the plane to England.
053PGenJ&MJUN03	JAM (5;3.12) MEG (7;8.6), MOT, GRA, BEC	At Grandma's house, Ware, England. Playing the game 'Guess who?' with British aunt.
054PGenJ&MJUN03	JAM (5;3.13) MEG (7;8.7), MOT, JAK, MAX	At cousins' house, Twickenham, England. Playing the game 'Kerplunk' in the bedroom with British cousins.
055PGenJ&MJUN03	JAM (5;3.13) MEG (7;8.7), MOT, JAK, MAX	At cousins' house, Twickenham, England. Playing 'Snakes and ladders' with British cousins.
056FPenJJUN03	JAM (5;3.21), MOT	At cousins' house, Twickenham, England. In kitchen, James playing with a brick garage, cars and a beyblade.

057CHenJ&MJUN03	JAM (5;3.24) MEG (7;8.18), MOT, GRA	At Grandma's house, Ware, England. James having a bath while Meggie chats.
058FPenJ&MJUN03	JAM (5;3.29) MEG (7;8.23), MOT, GRA	At Grandma's house, Ware, England. In the living room, looking at a comic and colouring in.
059TIptJJUL03	JAM (5;4.3), MOT, VIN	At Grandma's house, Ware, England. On telephone, James talking to his friend Vincent in Brazil.
060TIptJ&MJUL03	JAM (5;4.3) MEG (7;8.27), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
061CHenJ&MJUL03	JAM (5;4.10) MEG (7;9.4), MOT	At Grandad's holiday cottage in Newcastle, England. Lying in bed chatting to mother about the day.
062TIptJ&MJUL03	JAM (5;4.11) MEG (7;9.5), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
063PGenJ&MJUL03	JAM (5;4.16) MEG (7;9.10), MOT	At Grandma's house, Ware, England. Playing mini-snooker in the living room.
064TIptJ&MJUL03	JAM (5;4.16) MEG (7;9.10), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
065TIptJ&MJUL03	JAM (5;4.23) MEG (7;9.17), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
066PGenJ&MJUL03	JAM (5;4.26) MEG (7;9.20), MOT, JAK, MAX, BEC	At cousins' house, Twickenham, England. Playing word/sound bingo.
067CHenJJUL03	JAM (5;4.28), MOT	At Grandma's house, Ware, England. James talking to mother about the day spent at Chessington zoo.
068LAenJ&MJUL03	JAM (5;4.28) MEG (7;9.22), MOT, BEC	At Grandma's house, Ware, England. Meggie reading a story out loud to mother.
069TIptJ&MAUG03	JAM (5;5.0) MEG (7;9.24), MOT	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
070LAenMAUG03	MEG (7;9.27), MOT	At Grandma's house, Ware, England. Meggie reading the story 'Here I am, said Smedley' out loud to mother.
071TIptJ&MAUG03	JAM (5;5.7) MEG (7;10.1), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
072TIptJ&MAUG03	JAM (5;5.8) MEG (7;10.2), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father and other relatives in Brazil.

073MTenJ&MAUG03	JAM (5;5.14) MEG (7;10.8), MOT, GRA, WIL	At Grandma's house, Ware, England. Having dinner with Grandma and Uncle William.
074TIptJ&MAUG03	JAM (5;5.14) MEG (7;10.8), MOT, PAI	At Grandma's house, Ware, England. On telephone, siblings talking to their father in Brazil.
075TIptMAUG03	MEG (7;10.10), MOT, PAI	At Grandma's house, Ware, England.On telephone, Meggie talking to her father in Brazil.
076MTenJ&MAUG03	JAM (5;5.23) MEG (7;10.17), MOT, PAI	At home, Fortaleza, Brazil. Having lunch.
077CHenJSEP03	JAM (5;6.1), MOT	At home, Fortaleza, Brazil. James and his mother are looking at a photograph album of him as a baby.
078CHenJ&MSEP03	JAM (5;6.11) MEG (7;11.5), MOT, PAI	At home, Fortaleza, Brazil. In the bedroom, looking for toys and chatting about a book while eating.
079MTenJ&MSEP03	JAM (5;6.19) MEG (7;11.13), MOT, PAI	At home, Fortaleza, Brazil. Having breakfast before travelling to the mountains.
080MTenJ&MSEP03	JAM (5;6.26) MEG (7;11.20), MOT	At home, Fortaleza, Brazil. Having lunch.
081MTenJ&MOCT03	JAM (5;7.3) MEG (7;11.27), MOT, PAI	At the beach house in Redonda, Brazil.Eating spaghetti in the kitchen.
082CHenJOCT03	JAM (5;7.12), MOT	At home, Fortaleza, Brazil. Talking in bedroom about James' broken beyblade.
083MTenMOCT03	MEG (8;0.6), MOT	At home, Fortaleza, Brazil. Eating and chatting in the kitchen.
084MTenJ&MOCT03	JAM (5;7.18) MEG (8,0.12), MOT, PAI	At home, Fortaleza, Brazil. Having breakfast in the kitchen.
085CHenJ&MOCT03	JAM (5;7.24) MEG (8;0.18), MOT, PAI	At the beach house in Redonda, Brazil. In the bedroom talking before going to sleep.
086MTenJNOV03	JAM (5;8.0), MOT	At home, Fortaleza, Brazil. Having breakfast in the kitchen.
087MTenJ&MNOV03	JAM (5;8.0) MEG (8;0.25), MOT	At home, Fortaleza, Brazil. Having breakfast in the kitchen.
088MTptJ&MNOV03	JAM (5;8.7) MEG (8;1.1), MOT, PAI	At the beach house in Redonda, Brazil. Eating sandwiches in the kitchen.
089MTenJ&MNOV03	JAM (5;8.16) MEG (8;1.10), MOT, PAI	At home, Fortaleza, Brazil. Having breakfast in the kitchen.
090MTenMNOV03	MEG (8;1.19), MOT	At home, Fortaleza, Brazil. Having lunch in the kitchen.
091CHenJ&MNOV03	JAM (5;8.25) MEG (8;1.19), MOT	At home, Fortaleza, Brazil. Looking at James' school work.
092FPenJ&MDEC03	JAM (5;9.6)	At home, Fortaleza, Brazil. Playing at pretending

	MEG (8;2.0), MOT	to be ghosts.
093TIenJ&MDEC03	JAM (5;9.24) MEG (8;2.18), MOT, GRA	At home, Fortaleza, Brazil. On the telephone, the siblings talking to their Grandma in England.
094TIenJ&MDEC03	JAM (5;9.24) MEG (8;2.18), MOT, GRD	At home, Fortaleza, Brazil. On the telephone, the siblings talking to their Granddad in England.
095TIenJMAR04	JAM (6;0.0), MOT, GRA	At home, Fortaleza, Brazil. On the telephone, James talking to his Grandma in England.
096CHenJ&MAPR04	JAM (6;1.27) MEG (8;6.21), MOT	At home, Fortaleza, Brazil. Talking about the pregnant guinea pig and then the toys being taken to England on their move there.
097MTptJ&MMAY04	JAM (6;2.5) MEG (8;6.30), MOT, PAI	At home, Fortaleza, Brazil. Having lunch.
098MTenJJUN04	JAM (6;3.7), MOT	At home, Fortaleza, Brazil. Having breakfast and chatting about what James was looking forward to doing in England.
099MTptJ&MJUN04	JAM (6;3.8) MEG (8;7.2), MOT	At home, Fortaleza, Brazil. Having lunch in the kitchen.
100TIptJ&MJUN04	JAM (6;3.18) MEG (8;7.12), MOT, PAI	At home, Ware, England. On telephone, siblings talking to their father in Brazil.
101PGptJ&MJUN04	JAM (6;3.21) MEG (8;7.15), MOT	At home, Ware, England. Playing mini football on a mini snooker table.
102TIptJ&MJUN04	JAM (6;3.28) MEG (8;7.22), MOT, PAI	At home, Ware, England. On telephone, siblings talking to their father in Brazil.
103PGptJ&MJUL04	JAM (6;4.5) MEG (8;7.29), MOT	At home, Ware, England. Playing the game 'Guess who?'.
104TIptJ&MJUL04	JAM (6;5.21) MEG (8;8.15), MOT	At home, Ware, England. On telephone, siblings talking to their father in Brazil.
105FPptJ&MJUL04	JAM (6;5,25) MEG (8;8.19), MOT	At home, Ware, England. Siblings washing golf balls in the sink in the bathroom. Mother mostly absent.
106TIptJ&MJUL04	JAM (6;5.25) MEG (8;8.19), MOT, PAI, VOV, AVO	At home, Ware, England. On telephone, siblings talking to their father and other relatives in Brazil.
107FPptJ&MAUG04	JAM (6;6.6) MEG (8;9.0), MOT	At home, Ware, England. Siblings making a train track out of lego.
108MTenJ&MAUG04	JAM (6;6.6) MEG (8;9.0), MOT	At home, Ware, England. Having lunch.
109TIptJ&MAUG04	JAM (6;6.9) MEG (8;9.3), MOT, PAI	At home, Ware, England. On telephone, siblings talking to their father in Brazil.
110TIptJ&MAUG04	JAM (6;6.21) MEG (8;9.15), MOT, PAI, VOV AVO	At home, Ware, England. On telephone, siblings talking to their father and other relatives in Brazil.
111MTenJ&MAUG04	JAM (6;6.28) MEG (8;9.22), MOT, PAI	At home, Ware, England. Having dinner at the Clement Street house for the first time.

112LAenJ&MOCT04	JAM (6;7.6) MEG (9;0.0), MOT	At home, Ware, England. Meggie reading the book 'Rosie and the robbers' to her mother.
113LAenJ&MOCT04	JAM (6;7.10) MEG (9;0.4), MOT	At home, Ware, England. James reading the book 'Budgie the little helicopter' to his mother.
114TIptJ&MOCT04	JAM (6;7.13) MEG (9;0.7), MOT, VIN	At home, Ware, England. On telephone, siblings talking to their friend Vincent in Brazil.
115TIptJ&MOCT04	JAM (6;7.21) MEG (9;0.22), MOT, VOV, SAR	At home, Ware, England. On telephone, siblings talking to relatives in Brazil.
116TIptJ&MNOV04	JAM (6;8.5) MEG (9;0.30), MOT, VOV	At home, Ware, England. On telephone, siblings talking to their Brazilian grandfather in Brazil.
117TIptJ&MNOV04	JAM (6;8.27) MEG (9;1.21), MOT, PAI, SAR, VOV, JAN	At home, Ware, England. On telephone, siblings talking to their father and other relatives in Brazil.
118TIptJ&MDEC04	JAM (6;9.23) MEG (9;2.17), MOT, SAR, VOV	At home, Ware, England. On telephone, siblings talking to their relatives in Brazil.
119TIptJDEC04	JAM (6;9.19), MOT, PAI, VIN	At home, Ware, England. On telephone, James talking to his father and friend in Brazil.

Appendix B. Further details on the transcription and coding of the LOBILL Corpus

The description that follows is divided into three basic sections which reflect the three major components of CHAT: the file headers, the main tier, and the dependent tiers.

B1. File headers

The majority of file headers occur at the beginning of the transcript and contain information about the participants and the setting. They all start with the @ sign and are immediately followed by the header name. Most header lines will contain an 'entry' which gives additional information; this is typed after a colon and a tab, for example, **@Age of JAM: 5;7.15 .** A header is never followed by a punctuation mark.

The CHAT manual contains a set of headers which have been considered useful by researchers but it is also possible to create your own. However, there are a small number of headers which are obligatory; without these CLAN cannot perform its analyses correctly. After describing this group of obligatory headers, two optional initial headers will be mentioned. A third group of headers relating specifically to participants will then be shown, following which two other groups of headers will be described: 'constant' headers and 'changeable' headers.

B1.2 Obligatory headers

There are five headers which must be inserted into any transcript. Here, they are presented in the order in which they must occur:

B1.2.1 **@Begin**

This is a 'bare' header and as such is not followed by any information. It always occurs at the very beginning of a transcript and is used to ensure that no part of the transcription has been cut off or deleted by accident.

B1.2.2 **@Languages:**

Always the second line of the file, this header gives the languages of the transcript. If there is more than one language, that which predominates in terms of quantity comes first. Therefore, in the LOBILL Corpus, this header can appear as either **@Languages: eng, por** or **@Languages: por, eng .** Consisting of three letters, the language codes for over 40 languages are listed in the CHAT manual (pp 25-26).

B1.2.3 **@Participants:**

This third header contains a list of all the participants of the interaction. Each entry comprises three elements: 1) the speaker code; 2) the name of the speaker; 3) the role of the speaker. The speaker code is a combination of three capital letters which must be unique to each speaker (for automatic analysis purposes). It can be based on the name of the speaker or their role, for example, **MEG** (Meggie) or **CHI** (child). The second element gives the name of the speaker and will need an underline if there are two or more words, for example, **James_Lonngren**. For the final element, which provides information about the role of the speaker, there is a fixed set of roles from which you should choose. This specified set is used by the CLAN command CHECK which checks the transcript for errors. An example of a completed Participants header in the LOBILL Corpus is as follows:

@Participants: JAM James Target_Child, MOT Cathy Mother, PAI Alexandre Father

Although the recommended CHAT three-letter code for the role of father is **FAT**, a different code was chosen for the LOBILL Corpus: the code **PAI**, the word for father in Portuguese, was chosen to reflect the linguistic role of this particular speaker whose maternal language was Portuguese. For the same reason, the Brazilian grandparents were identified by the codes **VOV** (abbreviation of Vovô which means 'Grandad') and **AVO** (from Avó which means 'Grandmother'). All of the speaker codes used in the LOBILL Corpus can be found in Annex ??.

B1.2.4 **@ID:**

This fourth obligatory header contains further details about each speaker. There are nine possible fields, shown as follows:

@ID: language(s) | corpus | code | age | sex | group | SES | role | education |

It is not necessary to fill in all the fields and here are two examples of completed ID headers from the LOBILL Corpus:

@ID: por eng | lobill | MEG | 7;9.3 | female | | | Target_child | |

@ID: por eng | lobill | PAI | | | | Father | |

For bilingual speakers, two or more language codes can be inserted and in the LOBILL Corpus the order of the language codes is determined simply by the language spoken in the country of the speaker's birth.

B1.2.5 **@End**

This fifth obligatory header occurs at the very end of a transcript and has the same function as **@Begin**; it ensures that the transcript is complete and has not been inadvertently truncated (in the process of copying, for example).

B1.3 Optional initial headers

Apart from the five obligatory headers, there are two other headers which, if used, need to be placed in the beginning section of a transcript. They are mentioned below.

B1.3.1 **@Options:**

This header needs to come after the Participants header and specifies the checking rules which need to be suspended for certain files. For example, for a transcript containing sign language, words entirely in capital letters need to be allowed. By default the CHECK command normally restricts the use of capitals within words but, by inserting the following header **@Options: sign**, this action is suspended. Seven other options for suspension can be found in the CHAT manual (2011:28).

B1.3.2 **@Media:** The second of the two optional headers which must come at the start of the transcription is **@Media:** . Placed immediately after the **@ID:** header, the information in this header directs CLAN to the sound or video file linked to the transcript and allows it to be played. A completed Media header in the LOBILL Corpus looks like this:

@Media: 089enJ&MNOV03, audio, unlinked

The first field identifies the name of the file (the original extension, **.wav** in this case, should be omitted). The second field tells CLAN if it is an audio or video file while the last field informs whether it is linked or not to the transcript.

B1.4 Participant-specific headers

Although most information related to each participant can be found in the **@ID** header, the following three headers provide further information about the speaker: **@Birth of #: , @Birthplace of #: , @L1 of #: .**If used, they would come after the **@Media** header. For corpora with many different speakers or with speakers who may only appear in one or two files, this information could be important in assisting the researcher when selecting files for analysis and helping keep track of these variables when analysing results. In the LOBILL Corpus, however, the number of main speakers is very reduced and they represent a constant throughout the corpus. As CHILDES requires general information about the corpus and details about its main speakers to be included in an introductory document, the inclusion of these headers in each transcript would amount to unnecessary repetition. Therefore, the decision was made not to use the three headers mentioned above.

B1.5 Constant headers

The fourth group of headers discussed in this section are all optional and their function is to provide further information about the file in which they are inserted. Following the participant-specific headers, they are 'constant' in the sense that the information contained in them applies to the whole file and not just to one section. Currently 12 different constant headers can be found in the CHAT manual (MacWhinney, 2011, 30-31). Of these, seven were not used in the LOBILL Corpus, either because they were considered irrelevant, such as **@Room Layout**: or **@Recording Quality**:, or because the information they provided applied to the whole corpus and not just specific files, such as **@Transcriber**: or **@Transcription**: (this last one contains information about whether the transcription is partial, full, coarse, checked etc). As mentioned above, information applying to the whole corpus is necessarily included in an introductory document, thereby making it unnecessary in the case of the LOBILL Corpus to insert details such as the name of the transcriber or the nature of the transcription in each file (since they remain the same throughout).

Five of the constant headers which were considered relevant for insertion in the LOBILL Corpus are described below in alphabetical order (they may be inserted in any order after the participant-specific headers). A further header, **@Filename:**, which requires special discussion will also be described in this section.

B1.5.1 **@Interaction Type**:

Entries in this header may include **phonecall telechat, meeting, work, medical, classroom, family** etc. By adding this information to each file in the LOBILL Corpus, it is then possible to automatically select certain groups of files for analysis. For example, dinner-time interactions could be separately analysed from phone call interactions, or from interactions involving play or reading activities. Results could then be compared to see the possible relationship between the type of interaction and the code-switching practices found.

B1.5.2 **@Location:**

Containing information about the city, state and country of the interaction, an example of a completed header for the LOBILL Corpus is as follows: **@Location: Fortaleza, CE, Brazil**. As recordings were carried out in different locations and countries, this variable is important to be considered when analysing the data.

B1.5.3 **@Tape Location**:

In order for the researcher to locate the original recording, this header provides information about the tape number, side and footage, a completed example being as follows: **@Tape Location: Tape 1 Side B.**

B1.5.4 **@Time Duration:**

In the CHAT manual this entry states the time the recording started and ended, for example, **@Time Duration:12:30-13:30**. For many of the recordings carried out for the LOBILL Corpus these times were not noted and therefore this header states only the length of the recording, as the following example shows:

@Time Duration: 00:00:00-00:38:26

Specific information about the time of day of the recording can generally be found in the **@Situation**: header of each transcript (see B1.6). The lack of precise times in each transcript did not hamper the analysis of the results.

B1.5.5 **@Warning:**

This header provides observations about any aspect of a file which may restrict its use for specific analysis purposes. Such restrictions might relate to certain phenomena not being transcribed, such as retracings or overlaps, or it may be that

the file has not been double-checked for accuracy. In the case of the LOBILL Corpus, two warning headers feature in a few of the files. They are the following:

@Warning: Meggie was fully aware of being recorded and this may have influenced her use of Portuguese in the interaction

@Warning: This recording is of a telephone conversation where the Father's turns are not transcribed. His turns are perceived and may be inaccurately placed.

In the first case, MEG was being 'interviewed' by her mother in a set up which was considered more formal than a natural conversation. As such it was noted that this formality appeared to affect the linguistic dynamics of the conversation. Equally informative is the second warning which applies to all the transcripts of phone calls, but not Skype calls, in the Corpus. For other researchers wishing to examine the LOBILL Corpus, it is vital that they are warned about the peculiarities of some of the files and take these into account when analysing the data.

B1.5.6 **@Filename**:

This particular header deserves special mention. In the current version of the on-line CHAT manual (2014b), this particular header no longer appears in the section on headers. Constantly updated since its last printed publication in 2000, the on-line CHAT manual contains changes which have been made in the transcription system in order to improve the possibilities of analysis through the CLAN tool. In the 2004 version, which guided the initial construction of the LOBILL Corpus, the **@Filename: header** could be found under the 'constant' headers section of the manual. Despite being an optional header, it was considered a useful header to have as it provided the name of the file within the actual transcript. This duplication ensured that accidental renaming of files would be easily noticed: the name of the file found at the top of the CLAN editor window could be checked against the **@Filename:** header. to ensure that they were the same. For this reason, it was decided that this header would continue to be used in the LOBILL Corpus.

Decisions regarding the naming of each file in the corpus were discussed in detail in 3.1.5. As the entry for this header is simply the name of the file, a typical example would be as follows: **@Filename: 003enMSEP01.cha**

No further discussion of this particular header is needed here and I will now move on to describe the 'changeable' headers inserted in the LOBILL Corpus .

B1.6 Changeable headers

In contrast to the 'constant' headers described above, which refer to the transcript as a whole, 'changeable' headers may refer to specific sections of the transcript and can thus be inserted at appropriate points. However, if it is the case that a particular header applies to the entire transcript, it will come after the last 'constant' header.

Of the 12 changeable headers described in the CHAT manual (ibid:31-34) three were considered relevant for the LOBILL Corpus and they are described as follows.

B1.6.1 **@Date:**

In the case of transcripts which contain material recorded over more than one day, a new 'date' header would be inserted at each appropriate point in the transcript. As none of the transcripts in the LOBILL Corpus contain a recording which carried over into another day, the date does not change within each transcript. This means that this header will only need to occur once and be placed at the beginning of the transcript after the constant headers. An example of a completed 'date' header is as follows:

@Date: 04-OCT-2003

B1.6.2 **@Situation:**

This header contains general information about the setting and the activites in which the participants are involved. Two examples from the LOBILL Corpus are the following:

@Situation: JAM and MEG are eating spaghetti in the kitchen

@Situation: JAM and MEG are playing with lego in the bedroom

Once again, in the LOBILL Corpus, the nature of the short recordings means that the situation does not often change and for this reason, this header can be placed under the constant headers. In other corpora, where there is a change in setting or activity during a recording, this header should be inserted again at the relevant point.

B.1.6.3 **@Comment:**

This third changeable header is an all-purpose header which can have as its entry various types of relevant information. If the comment to be made applies to the whole transcript, it comes after the constant headers. If it refers to a specific section of the transcript it occurs at the relevant point. An example of a **@Comment** header occurring within a transcript in the LOBILL Corpus is the following:

@Comment: William is talking to Grandma as MEG and her Mother start talking

The transcript in question is a recording of a dinner-time conversation where the participants include the two children, their mother, grandmother and uncle (William). At a certain point in the dialogue, William begins talking to the grandmother while the mother engages in conversation with her daughter, MEG. By inserting the comment, the researcher is better able to follow the ensuing dialogue.

If the comment to be made refers specifically to one particular utterance, this should be done by using a different type of coding (see section 3.2.3.2). Having described the first component of the CHAT transcription system, the file headers, and how they will be used in the LOBILL Corpus, the description will now move on to how speech is transcribed on the 'main lines'.

B2. Main lines

It is on the main line that speaker utterances are transcribed. Each main line must begin with an asterisk (*), immediately followed by a three-letter speaker code, a colon and a tab. The utterance itself can then be typed in, ending with one of the basic terminators: a full stop, an exclamation mark or a question mark. Variants on these basic terminators will be shown in section B2.2.2. A typical utterance would therefore look like this:

***MEG: give me that please.**

Note that transcribed utterances do not normally begin with capital letters. Typical exceptions would be the personal pronoun 'I' and proper nouns occurring in initial position.

By following these minimum conventions, it is already possible to perform several types of analyses on the main line with the CLAN tools. However, CHAT

offers an array of further transcription conventions which provide a much richer representation of spoken discourse and this consequently allows for more complex analyses to be performed on the data. Essentially it is the type of analyses that a researcher wishes to perform on a particular corpus that will determine the range of conventions that need to be used when transcribing the data on the main line. Before describing those used in the LOBILL Corpus, it is necessary to discuss certain issues related to the transcription of 'words' and define what is meant by an 'utterance'.

B2.1 Transcribing words

One of the goals of the CHAT system is to 'maximize systematicity and minimize inconsistency' (MacWhinney, 2014b:36). Inconsistencies in the transcription of words will lead to inconsistencies in the results of computational analyses at word level. In order to maximise the potential of the CLAN tools for lexical and syntactic analyses, it is therefore crucial that lexical items are represented clearly and consistently throughout a particular corpus and across corpora. The CHAT transcription system provides specific ways for transcribing all types of words: common words, compounds, acronyms, numbers, titles, shortenings, assimilations and exclamations among others. These conventions will be discussed and illustrated below with examples from the LOBILL Corpus.

B2.1.1. Common words

To achieve standardized spellings of common English words, CHAT uses Webster's Third New International Dictionary as a reference. Providing standard American English spellings of words, it is evident that there are differences to be found between the spelling of certain words in this variety of English and that of standard British English. Common examples include 'color' and 'Mommy' in American English as opposed to 'colour' and 'Mummy' in British English. On examination of the British English corpora in the CHILDES data base, standard British English spellings were found. Thus it is clear that the system caters for the transcription of different varieties of English. It appears that the key here is to ensure a certain standardisation within and across a variety. To achieve this it was decided to use two on-line dictionaries as references: the Oxford English Dictionary (Second Edition) (http://www.oed.com) and the Michaelis Moderno Dicionario da Lingua Portuguesa

(http://michaelis.uol.com.br). This information would be included in the introductory document submitted to the CHILDES data base about the LOBILL Corpus.

B2.1.2 Compound nouns

CHAT offers guidelines for the transcription of a variety of compound types: true compounds, collocations and linkages. For purposes of analysis, it is recommended that true compounds which are normally written as one word be written using the + symbol to link the two, or more, parts of the compound. Therefore, **fireman** becomes **fire+man** and **t-shirt** becomes **tee+shirt**. It is important to note that dashes or hyphens in compounds (like in the latter example) should be replaced by the + symbol: in CHAT the dash is used to indicate suffixation on an optional morphology line below the main line.

When it comes to compounds or collocations normally written as two words, it is up to the transcriber to decide whether to treat them as separate words or to use the + symbol to link them up. In the LOBILL Corpus, it was decided that the + symbol would be used in these types of compounds and strong collocations, as in the following examples: **fire+engine, fire+brigade, air+conditioning, teddy+bear, swimming+pool**.

A third category of 'compounds', referred to in the manual as 'linkages' (2014b: 45) are collocational phrasal combinations which are normally written as separate words in standard orthography. As the following examples show, it is recommended that they be written using underscores to indicate their collocational relationship: **Thomas_the_tank_engine, Mister_Men, Holland_and_Barratt, British_Water_Ways, Lady_and_the_Tramp, Jingle_Bells, The_University_of_Hertfordshire**. Songs, book titles, films, places are typical examples of when the underscore can be used.

B2.1.3 Acronyms

A further use of the underscore in the CHAT system is when transcribing acronyms. Although common acronyms such as tv and dvd should be written as word forms and without capital letters, proper nouns should be written as follows: U_K, B_T, B_H_S. If the acronym is prounounced as a word form and not spelled out in letters, it should be written as a word beginning with a capital letter. For example, HESS (Hertfordshire Educational Supply Service) would be transcribed as Hess and the

LOBILL Corpus would become Lobill_Corpus. Full stops must not be used within acronyms as they can only act as utterance terminators.

B2.1.4 Letters

Where a speaker is spelling out a word, letters need to be transcribed by adding **@l** after each letter. An example sequence would be the following:

***MOT: <Meggie # how do you spell ball[“]>[@en]?**
***MEG: <b@l a@l l@l l@l>[@en].**

If *MEG were to have spelt it out in Portuguese, this would be indicated by swapping the **[@en]** code for **[@pt].**

B2.1.5 Titles

Instead of using abbreviations, titles need to be written in full. Therefore, for example, **Mr** must be written **Mister** and **Dr** should be written **Doctor**.

B2.1.6 Numbers

As with titles, numbers need to be written out in words. So **1998** would be written as **nineteen ninety eight** and **23.5 %** needs to be written as **twenty three point five percent**. If the transcriber wishes to treat a number sequence as a compound, the underscore can be used: **101 Dalmations** could be written **One_hundred_and_one_Dalmations.**

B2.1.7 Kinship forms

As expected, there are some differences in kinship address forms between different varieties of English. In the list provided in the CHAT manual (p47) there are two spellings of forms of address which cannot be found but which are used by the speakers in the LOBILL Corpus: **Mummy** and **Grandad**. Added to these particular spellings are all of the Portuguese forms of address used by the speakers, which include the following: **Papai** (Daddy), **Pai** (Dad), **Mamãe** (Mummy) **Mãe** (Mum), **Avó** (Grandfather) and **Avô** (Grandmother). As long as consistency is achieved across the corpus, these variants pose no problem when it comes to analysing the data.

B2.1.8 Exclamations and Communicators

Spoken discourse is punctuated with exclamations and communicators which can greatly vary in terms of their phonological form. To achieve maximum consistency, CHAT provides a list of spellings for these forms which should be used even if the phonological form uttered is slightly different (see pp. 49-50). Approximation is considered better practice than creating new spellings. Those chosen from the list for the LOBILL Corpus are shown below:

Table B1. Spellings and meanings of the exclamations and communicators used in the LOBILL Corpus

Exclamations

Expression	Meaning
ah	relief
ahhah	discovery
aw	solidarity
haha	amusement
mmm	tasty
ow	pain
sh	silence
ugh	disgust
uhoh	trouble
wow	amazement
yea	a cheer

Communicators

Marker	Function
ahem	ready to speak
er/um	pause
huh	questioning
hmm/mmm	thinking/waiting
hmm/huh?	questioning
uhhuh/mhmm	yes
uhuh	no

It was necessary to add only three more exclamations to cater for the variation found in the bilingual corpus: **hey!** (in English) or **ei!** (in Portuguese) to express indignation and **ai!** to express pain in Portuguese.

B2.1.9 Shortenings

Spoken discourse is characterized by shortened forms such as **(be)cause** or **(re)member**. In order to facilitate the automatic analysis of transcripts, it is important to transcribe these shortenings consistently and CHAT provides a simple way of doing this. As shown in the two examples above, the parts of the word which are not pronounced can be enclosed in brackets. When using the CLAN commands, the researcher is able to choose whether they wish the analysis to be done on the complete form or whether they want the part in brackets to be ignored. This is achieved by using the +r option (see 3.3.3).

Further examples of typical shortenings used by the speakers in the LOBILL Corpus include **(a)n(d), Gran(d)ma, (un)til, (Mum)my, (es)tá, p(a)ra.**

B2.1.10 Assimilations

Spoken forms which involve the assimilation of two or more words (often involving auxilaries, infinitives or pronouns), can also be transcribed in such a way so as to allow the researcher to perform analyses on the assimilations themselves (monomorphemic forms) or the complete forms. This is done by writing the complete form in square brackets after the assimilated form. Thus, in the LOBILL Corpus we will frequently see the following forms: **gimme[: give me], lemme[: let me], gonna[: going to],** and **wanna[: want to]** . A list of other assimilated forms can be found on page 49 of the CHAT manual.

B2.1.11 Dialectal variations

The notation method of square brackets can also be used to include standard spellings of forms which vary from the norm in terms of phonology. For example, when MEG pronounces 'then' as 'den', this can be transcribed in the following way:

***MEG: I fell over and den[: then] I cried.**

This form of notation will allow retrieval of all cases of 'then'.

Another method for transcribing dialectal variations is the use of full phonological transcription for the utterance. For researchers wishing to analyse the phonological features of spoken discourse this method would clearly be more useful. For the present research, however, this is not the case and the simple bracket method is used.

B2.1.12 Baby talk

For the transcription of onomatopoeic words and diminutives such as **wowwow** and **doggy**, a list is provided of some common forms on pages 52-3 of the CHAT manual. The general recommendation is that diminutives should be written with **ie** at the end (except for very common forms such as **doggy, kitty, potty, tummy, dolly**). Examples would include **ballie, forkie, horsie**. Whatever the baby form may be, where the meaning of the word is unclear it is possible to indicate a referent on the main line or underneath the utterance on the dependent tiers (see section 3.2.3.2 for the latter). Thus, in the LOBILL Corpus when the meaning of the word is not obvious, an explanatory referent can be found in brackets after an equals sign, as shown in the following example:

***JAM: where is my bibi [= dummy]?**

Specific baby forms which are used very frequently in a corpus may be compiled in a list and included in the introductory document: this serves to help other researchers wishing to analyse a corpus they are unfamiliar with.

B2.1.13 Prosody within words

Before moving on to the discussion of the transcription of utterances, as opposed to words, it is relevant here to briefly descibe how prosody within words can be marked. To indicate an elongated syllable, a colon can be used:

***MEG: is that a ti:ger?**

To mark a pause between syllables the ^ symbol can be used:

***MEG: have you ever seen frozen water in Eng^land?**

For stressed syllables, a triple forward slash can be inserted immediately before the syllable in question:

***MEG: I've al///ready done it!**

As has already been mentioned, despite these coding options being available, it is up to the researcher to choose the extent to which word-level detail, such as prosody within words, will be coded in a particular corpus. This is also true of coding at the utterance level, as will be seen in the next section.

B2.2 Transcribing utterances

In the CHAT transcription system each utterance must be transcribed on a separate main line. Therefore, as the following excerpt shows, although MEG maintains the floor, each utterance is treated as a different turn and transcribed on different main lines, one after the other.

***MOT: Meggie your[//] it's your turn.**
***MEG: whose turn is it after me?**
***MEG: me.**
***MEG: so, sorry you have to pick up four.**
***MOT: well!**

Each main line must end with one of the utterance terminators (a fullstop, question mark or exclamation mark). Note that commas can be used within utterances to mark pauses or syntactic junctures.

A single main line utterance may continue for more than one computer line and will look like this in the CHAT editor:

***MOT:** **do it again because otherwise it's gonna[: going to] be unfair.**

There is no need to add the tab space as it automatically goes on to the next line in the correct place.

Delimiting utterances can be a difficult task especially when the discourse to be transcribed contains repetitions. The manner in which words are grouped together and the way repetitions are treated by the transcriber will ultimately affect the results of later analyses, especially those calculating the MLU (Mean Length of Utterance). The next section will discuss issues relating to the transcription of repetitions and then other CHAT transcription conventions used on the main line will be illustrated.

2.2.1 Repetitions

The following excerpt from the LOBILL Corpus illustrates three different ways of transcribing repetitions. The discourse is between the mother and her daughter who are beginning a card game of Uno. The utterances have been numbered for ease of reference and the numbers do not form part of the transcription.

1. ***MEG:** **one one, two two, three three, four four, five five, six six, seven seven.**
2. ***MEG:** **okay.**
3. ***MOT:** **who's going to start?**
4. ***MEG:** **you because I shuffled[*] them.**
5. ***MEG:** **no, me me me.**
6. ***MOT:** **no no no no no.**
7. ***MEG:** **yes yes yes.**
8. ***MOT:** **no.**
9. ***MOT:** **<you sh(uffled)>[/] <you shu(ffled)>[//] you dealt.**

In MEG's first utterance (1.) she is dealing out seven cards to herself and her mother, counting them out as she deals them into two piles in an alternating sequence. The repetition of each number is propositional and accompanies the action of dealing. The use of the comma between each number indicates the prosodic grouping of the repetitions but the pauses are not sufficient to warrant separate utterances.

Once the cards are dealt there is then an exchange to decide who is going to go first. In line 5. MEG emphasises the fact that she wants to go first by repeating **me** three times in quick sucession, the lack of commas indicating that there are no pauses between the repetitions. Following this pattern the mother then repeats **no** five times without pauses (line 6.) with MEG then contesting this denial with three quick repetitions of **yes** (line 7.). The transcription of these repetitions without commas and all contained on one line indicates that all of them were intentional and said in quick succession.

The above cases contrast with the repetition that can be found in line 9. Here the mother wants to say that she should go first as MEG had dealt the cards. However, before eventually managing to say **you dealt**, the mother makes an incomplete repetition of **you shuffled,** influenced by MEG's mistaken use of this verb in line 4. (errors are indicated by the asterisk). To indicate that this was an involuntary repetition, the words concerned (the first time they appear) are put between angled brackets and then followed by square brackets containing a forward slash **[/]**. As the mother then goes on to reformulate her utterance by changing **you shuffled** to **you dealt**, the former is enclosed in angled brackets and followed by square brackets containing two forward slashes [//]. This symbol is used to indicate that the ensuing reformulation incorporates words from the material within the brackets (in this case the word **you**). A complete reformulation would be preceded by square brackets containing three forward slashes **[///]**. The use of such coding allows researchers to choose whether they wish to include or exclude such material from certain analyses. For example, although the MLU programme, by default, automatically excludes this material from its analyses, a switch (+r6) can be used to turn off this default. This differs from the WDLEN programme (used to analyse the LOBILL Corpus) which by default includes such material in its analysis. In this case the same switch can be used to effect its exclusion (see pp.124-125 of the CLAN manual, 2014c).

It is evident from the above discussion that the method of transcription of repeated words will depend on their discursive function. And it may often be the case that each repetition needs to be treated as a separate utterance, as in the final example shown below:

***JAM: mummy.**

***JAM: mummy.**
***JAM: mummy!**
***MOT: what +!?**

Here, JAM is trying to get his mother's attention without much success. After getting no response the first time, he then tries again. However, it is only after saying it the third time with more forcefulness that his mother finally answers him. By transcribing the repetitions this way, it is being shown that each **mummy** is an independent utterance and being treated as a separate turn.

Whatever decisions are made concerning the transcription of repetitions, again the goal needs to be consistency. This applies to all the annotation used in a particular corpus, including the other optional CHAT codes which can be used to transcribe on the main line, discussed as follows.

B2.2.2 Special utterance terminators and linkers

Allthough it is a formal requirement of CHAT that each utterance must finish with either a fullstop, question mark or exclamation mark, they may be immediately preceeded by a combination of other punctuation marks which serve to mark special discourse features. The special terminators and linkers (which occur at the beginning of utterances) used in the LOBILL Corpus are shown below:

Table B2. The special terminators and linkers used in the LOBILL Corpus

Special terminators	Meaning
+...	trailing off
+..?	trailing off of a question
+!?	question with exclamation
+//.	self-interruption
+//?	self-interruption of a question
+/.	interruption (by another speaker)
+/?	interruption of a question (by another speaker)
+"/.	quotation follows
Linkers	
+,	self-completion
++	completion (by another speaker)

+”	start of quote
+^	quick uptake
+<	overlaps previous utterance

Some of the terminators are often used in conjunction with a ‘linker’ which indicates how the previous utterance is linked to the following one, or ones. In the examples below we can see how terminators and linkers can work together to give greater discourse detail:

***JAM: that oh I want that +/.**
***MOT: +< you’re not eating that!**
***JAM: +, sweetie.**

Here JAM is interrupted by his mother (indicated by the terminator +/**.**). To show that MOT’s utterance overlaps JAM’s a linker is used (+<) and then to show that JAM completes his utterance despite the interruption another linker (+**,**) is used. By linking the two parts of JAM’s utterance in this way, CLAN is able to perform MLU analyses on the complete utterance.

In the following example, when MEG doesn’t finish her question (indicated by +**...**), MOT offers a completion (indicated by the linker ++).

***MEG: have you got a +..?**
***MOT: ++ biscuit?**

When it comes to quoted discourse a specific linker (+**”**) must be used with the preceeding terminator, as can be seen in this example:

***JAM: é assim ô +”/.**
***JAM: +” there’s too much butter on those trays.**
***JAM: then Manuel say like this, look, que[“]?**

In the LOBILL Corpus the coding of quotations proved to be very important as the beginning of quoted discourse often marked the location of a code-switch (as can be seen in the example above). When analysing the output of a particular analysis (for example, a list of concordances) the presence of the quotation codes (+”/. and +”) will allow the researcher to investigate the link between code-switching and the quoting of direct speech.

B2.2.3 Paralinguistic information

Although the coding of paralinguistic events is optional it provides the reader (and the CLAN tools) with much richer material to analyse. These events may or may not be accompanied by speech and CHAT provides options for coding on the main line for both cases, as will be seen below.

Firstly, for sounds or actions which occur without speech, the prefix &= is used, followed by the sound or action. Examples from the LOBILL Corpus include the following: **&=laughs. , &=coughs. , &=sighs. , &=belches.** . For the sake of consistency, standardized spellings for English are offered in the manual (p. 61). For those events which are accompanied by vocalization and/or actions, the following format can be used: **&=imit:motorbike. , &=imit:dog. ,** (where **imit** is an abbreviation of **imitation**) **&=points:car. , &=runs:door. , &=head:yes. , &=eats. , &=drinks.** . This coding allows for the inclusion of transitive objects (placed after a colon).

In the second case, where the transcriber wishes to code paralinguistic information occurring with speech, CHAT provides three options for coding on the main line. Those used in the LOBILL Corpus are exemplified in the following discussion.

In these first two examples, we can see exactly the same utterance coded slightly differently.

***MEG: I don't want it there[=! shouts]!**

***MEG: <I don't want it there>[=! shouts]!**

In the first one the code in square brackets indicates that MEG raised her voice when she said the word **there**. To indicate that MEG shouted out the whole utterance, scoped symbols (the angled brackets) would need to be placed around the entire utterance, as can be seen in the second example.

Where the paralinguistic information serves to specify references, the coding is very similar to that shown above, as can be seen here:

***PAI: that's mine[= cushion]!**

However, in this case there is no need for an exclamation mark after the equals sign.

If the information that needs to be included is more detailed and goes beyond a few words, the recommendation is to avoid including it on the main line so as not to impede the legibility of the utterances. More extensive comments can be placed underneath the main line on the dependent tier, as will be shown in section 3.3.3.

B2.2.4 Alternative or uncertain transcription

Where the transcriber is uncertain about what a speaker actually says, there are three ways to indicate this on the main line, depending on the degree of uncertainty. When there is no attempt at a best guess, individual words can be replaced by **xx** (two crosses) and groups of words or entire utterances by **xxx** (three crosses). If there is an attempt to work out what is being said individual words can be followed by **[?]** and two or more words can be enclosed within angled brackets and then followed by **[?]**. If the transcriber wishes to provide a possible alternative for their first choice of transcription the following code can be used to give the alternative: **[=? text]**. The following examples demonstrate these three different forms of coding:

***MEG: if she's like this, her womb stretched[?] xxx they must be close to coming out.**

In this example, MEG is talking about her pregnant guinea pig and not all of what she says is clear: a guess is made at the word stretched (indicated by the **[?]** and the three crosses (**xxx**) represent a sequence of words which proved impossible to transcribe. In the next example, the transcriber is unsure whether MEG says **get** or **pick** and provides the alternative in brackets:

***MEG: I know but I'm the one who goes with Sara to get[=? pick] the grass and everything.**

As mentioned above, by using the angled brackets, the scope of the coding increases to include all the material within the brackets, as in the following example:

***MEG: the[//] <more times>[?] I get it than you.**

Here, the two words **more times**, represent a best guess and are therefore enclosed in the scoped symbols.

B2.2.5 Pauses

Unfilled pauses can be coded in as little or as much detail as required by the specific research aims which underlie the construction of a corpus. In the LOBILL Corpus, pauses are coded in the following way:

***MEG: you're like (.) the same age as Biju.**

A fullstop within parentheses indicates a short pause, longer pauses being indicated with **(..)** and very long pauses with **(...)**. If required, The exact length of the pause in

seconds can replace the use of fullstops, so **(.)** could become **(2.5)**. However, this level of precision was not deemed necessary for the present corpus.

B2.2.6 Interposed words

Often it is the case that a listener may utter short words such as **yeah, mhmm** without interrupting a speaker's turn. To avoid unnecessarily having to break up a speaker's discourse and having to use linkers to link up the utterances (as seen earlier), it is possible to code such short interpolations within the utterance of the floor-holder. The following excerpt shows how this can be done (the underline itself does not form part of the coding, this is used here to simply indicate the location of the interpolation):

***MEG: well I don't wanna[: want to] give them away, I want to wait until I came[//][*] come to &*MOT:mmm visit I wanna[: want to] see them.**

For purposes of analysis, by using this coding, the interpolations are automatically excluded from any analyses carried out on the main speaker's utterances.

B3. Dependent tiers

After having described two major components of a CHAT transcript (the file headers and the main line), this section will discuss the third component of the CHAT system, the dependent tier. It is important to point out that, unlike the first two components, the use of the dependent tier is completely optional. Although there are no requirements for a researcher to provide information on the dependent tier, it is evident that by doing so the corpus becomes a much richer field for investigation purposes. CHAT provides several options for coding but also allows for the creation of novel codes for specific purposes.

All dependent tiers start with the percent symbol **%** which is then immediately followed by a three-letter code in lower-case and a colon. After a tab the metalinguistic information is then included. It is possible to insert as many dependent tiers as desired for each main line utterance, one below the other. The table below shows some of the more than twenty-five codes that can be found in the CHAT manual (pp. 78-85):

Table B3. 12 of the dependent tier codes available in CHAT

Code	Gloss	Use

%act:	action	describes the actions of the speaker/interlocutor
%add:	addressee	describes who is addressing who
%alt:	alternative transcription	provides an alternative transcription of the one on the main line
%com:	comment	used to make any types of comments
%eng:	English	provides a translation of the main line utterance
%err:	error	codes errors
%exp:	explanation	offers explanations, especially for identifying objects of people
%gls:	gloss	provides a 'translation' of a child's utterance into an adult form
%mor:	morphology	codes morphemic segments
%par:	paralinguistic	describes paralinguistic events such as coughing or crying
%pho:	phonology	gives a phonological transcription using the IPA[212]
%spa:	speech act	codes speech acts (for example, 'criticize' 'threaten')

Greater detail and examples of all the options can be found in the CHAT manual (ibid).

While accepting that the more coding there is, the richer the corpus will be for general investigation purposes, the time limitations of a research project mean that the extent of the coding must be determined by the specific research questions. For this reason, it was decided to make use of the following three dependent tier codes in the LOBILL Corpus: %add, %com and %err. Each of these was discussed and exemplified in Chapter 3 (see 3.2) and, as such, no more discussion is needed here.

The information in this appendix has served to provide a more comprehensive description of how the LOBILL Corpus has been transcribed and coded following the CHAT system.

[212] International Phonetic Alphabet

Appendix C. Non-word list (**@nonwords.cut**)[213].

Table B4. Complete list of non-words included in the **@nonwords.cut** file

mmm	ooh	uhh	ssss
err	ohhhh	ui	ssshh
mhmm	eh	uhmm	sssshhh
uh	deh	argh	sssshhh
erm	mm	aw	ststst
uhuh	ssh	grr	ugh
ha	hm	hoh	uhhu
ah	ahah	humph	umm
ahh	urgh	oo	urghh
hmm	uhhuh	pshh	wah
mmhm	aha	schh	wahh
woh	brr	schhh	whoh
woohh	er	shh	wooohhh
huh	heh	ssh+ssh	wey
choo	psshh	sshhhh	wiwiwiwiwiwi
ohh	tch	sssh	

[213] The format of the original 'cut' file is as a single list. Here it has been formatted into four columns to save space.