1 Introduction: Motivating a Unified Linguistic System

The literature on bilingualism takes it for granted that languages, grammars, and lexicons are countable entities; consequently, a bilingual is a person who has two languages, or two grammars, or two lexicons in their head.[1] This is why we can say things like: "Mary speaks three languages," "Joan speaks more languages than Chris does," etc. Moreover, languages are autonomous entities with clear boundaries. Let's call this the *separationist* framework or the *common-sense* view. According to this common-sense view, a bilingual has two grammatical systems in their head. This view is mainstream in linguistics and psycholinguistics. The quotation in (1) is an explicit statement of this assumption by a famous scholar of child bilingual acquisition.

(1)   In acquiring two languages from birth, children are undergoing a sort of "double" acquisition process in which two morphosyntactic systems are acquired as fundamentally separate and closed systems.

(De Houwer 2005: 43)

However, it is worth pointing out that this view is often implicit precisely because it is so commonsensical.

Within separationism, code-switching is defined as a "going back-and-forth" from one language to the other or as an "insertion" of items from one language into a discourse constructed out of ingredients from a different language. Consider the example in (2), a sentence planted in my facebook feed by some marketer:

(2)   *Antes de que se vaya*, thank President Obama for everything he's achieved. He's worked hard to protect and defend *nuestros terrenos, nuestro aire, nuestras aguas, nuestras comunidades, y nuestra madre tierra*. Add your name to our thank you letter today!
        ("antes de que se vaya" = "before he leaves")
        ("*nuestros terrenos, nuestro aire, nuestras aguas, nuestras comunidades, y nuestra madre tierra*" = our lands, our air, our waters, our communities and our mother earth")

1

For a separationist, this sentence follows the "back-and-forth" model: it is constructed by starting off in Spanish, switching then to English, going back to Spanish, and ending in English.

The following sentence exemplifies an "insertionist" style of code-switching:

(3)   She brought the *manguera.*
                        hose
                                Herring et al. 2010 (from the Miami corpus, CSBTP, Bangor)

The Spanish word *manguera* is "inserted" in an otherwise English monolingual discourse.

I argue that the separationist perspective on the study of bilingualism is hindering progress and we should take a different path to conceptualizing bilingualism. This monograph is a first step in that direction. I believe our default assumption should be that a single linguistic competence grows out of the faculty of language on the basis of whatever ingredients the environment supplies. There are no two lexicons or two PFs. Bilinguals have two systems of exteriorization that link the linguistic competence to the articulatory/perceptual systems. The I-language of bilinguals is not substantially different from monolinguals.

A brief consideration of post-Creole continua will help us understand the futility of understanding 'languages' as countable, discrete entities. As is well known, many speakers of Creole languages possess a spectrum of linguistic ingredients that range from the *basilect* – a form of expression that is very distant from the lexifier language – to the *acrolect* – a form of expression that is very close, or identical, to the lexifier language.

Figure 1.1 is an example of Guyanese Creole from Bell (1976). Depending on environmental conditions, the same person might say the simple transitive sentence 'I gave him one' in a variety of ways, from the Standard English to the basilect form, which constructs the same proposition with a very distinct lexical and grammatical structure.

Let me now invite you to try a thought experiment. Let's imagine a remote country, that we may call Twin Guyana, in which only the varieties 1 and 18 were spoken and accepted – and some inhabitants of Twin Guyana would use both. The naïve linguist who would alight on this island would probably conclude that these speakers are bilingual – and the "separationist" scholar would say that these people have two distinct grammatical systems and two distinct lexicons in their head. But in real Guyana, we don't find this separation, rather, speakers can use a variety of forms distributed along a continuum. The existence of the continuum makes it apparent that a separationist understanding

#					
1	aɪ	geɪv		hɪm	wʌn
2					wan
3	a			ɪm	
4				i:	
5		gɪv		hɪm	
6				ɪm	
7				i:	
8		dɪd	gɪv		
9		dɪ	gɪ		
10		dɪd			
11		dɪ	gi:		
12			gɪ	hi:	
13	mɪ				
14				i:	
15		bɪn			
16			gi:		
17				æm	
18					

Figure 1.1 Guyanese post-Creole continuum

of human language is misguided. In particular, the staggered distribution of the different forms is very revealing. That is: it is not the case that the different subject pronouns /ai/, /a/ and /mi/ each trigger a verbal form in a one-to-one fashion; rather, each is compatible with a range of verbal forms that partially overlap.

I think that a separationist scholar would have to posit that speakers of Guyanese Creole could have eighteen languages in their head (or a subset of those eighteen languages, depending on the social milieu in which they grew up). But that would be patently absurd. Instead, we need to change our assumptions and stop thinking of languages as countable entities if we want to understand post-Creole continua. Arguably, the insight supplied by post-Creole continua should be fully incorporated into our theory of I-language in general and bilingualism in particular.

But, how about monolinguals? Could we say something like, "sure, a Creole speaker's I-language is a continuum, but one could not meaningfully say that a monolingual grammar is a continuum too"? I disagree. As Tom Roeper (1999) and others have shown, the grammar of English includes so many nooks and crannies that, Roeper claims, English speakers all have multiple grammars in their head. I think the right way to conceptualize this claim is not as "many" grammars, but as "one" linguistic system with a range of options. We should operate under the assumption of an integrated I-language.

The problem highlighted by post-Creole continua is a familiar one in the natural sciences. Everything in nature is a continuum. Biologists tell us that the barriers between species are often arbitrary and what they find is a continuum of finches or starlings or felines which are very similar in adjacent regions (or regions that were adjacent in the past) but become more distinctive as geographical distance increases. So, if we take our individual I-language, as any other phenomenon of the natural world, to be a continuum, the interesting part is how to study this continuum in a formal, explicit way. I think a first step in this direction entails abandoning the separation hypothesis and taking bilinguals to have a single system of linguistic competence.

In linguistics research, one commonly sees arguments structured in the following manner: "theory T claims that there is a categorical distinction between A and B, and this distinction gives rise to a set of predictions. However, this distinction is a figment of the linguist's imagination because what we have instead is a continuum of phenomena between A and B. That is the end of the analysis: theory T is bad. Therefore, we should give up on categorical distinctions and be satisfied with the description of continua." I don't agree with this reasoning. Everything we have in nature is continua, but sciences have advanced precisely when they are able to develop conceptual frameworks to study these continua rigorously. Identifying the continua is not the end but the beginning of the analysis.

Let's turn now to code-switching, which is the main empirical focus of this monograph. One important feature of code-switching by early or deep bilinguals is that the items used in a code-switched sentence enter in "mixed" dependencies. Let me explain this with three examples:

(4)   *Basque*/Spanish
    a. *Ez   zen*   nadie   *etorri etxera*
        NEG AUX.PAST anyone come   home
        'No one came home'
    b. No vino   *inor*   a casa
        NEG come. PAST   anyone   to home
        'Nobody came home'

Vergara and López 2017: 270

(5)   German/*Spanish*
        Das   Buch, Hans *lo*   *hizo*   verkaufen.
        DEF.N   book   Hans CL.ACC.M   did   sell
        'Hans sold the book'

                                                        González-Vilbazo: p.c.

(6)   *Sranan*/Hindustani
        a. *plafond boro*   kare
           ceiling bore.through do
           'hit the ceiling'   Muysken 2000: 254
        Spanish/German
        b. *hizo*   nähen   das Hemd
           do.PAST   sew   the shirt
           'She/he sewed the shirt'

                                        González-Vilbazo and López 2012: 35

In example (4a), the Spanish word *nadie* is a negative polarity item, whose appearance in the sentence is licensed by the Basque sentential negation *ez*. These two words form a syntactic dependency even though they come from different languages. (4b) is the mirror image of (4a). (5) is an example of left dislocation in a Spanish/German code-switching variety. The dislocated constituent is the German *das Buch*, which is doubled by the Spanish clitic *lo*. They also form a ϕ-feature dependency, although German does not have an equivalent to Romance dislocations (it has some approximations with weak pronouns in lieu of clitics but the effects on the grammar are different). This dependency has consequences in semantic interpretation (*das Buch* is understood as a topic) and for phonetic representation (*das Buch* exhibits the peculiar intonation and caesura of dislocated constituents). Gender agreement between clitic and dislocate is particularly intriguing. German has a three-way gender system of masculine, feminine, and neuter, while Spanish only has masculine and feminine. Interestingly, the neuter DP *das Buch* is doubled with a Spanish clitic that is inflected for masculine gender – feminine would be unacceptable – which suggests that Spanish masculine and German neuter are somehow seen as having enough in common to establish a dependency.

Finally, consider now the examples in (6). They have in common that the predicate phrase consists of a light verb and a lexical verb as a complement. Example (6a) combines constituents from Sranan and Hindustani and (6b) combines Spanish and German. In (6a), the lexical verb and its complement are in Sranan and form a verb phrase with an idiomatic meaning. This verb phrase is the complement of the light verb *kare*, which is Hindustani. Interestingly, the verb and the object appear in OV order, which is the normative order in Hindustani, not in the VO order that is expected of Sranan. In this respect, (6a) contrasts with (6b). Although German normally requires OV when the VP is the complement of an auxiliary or modal, in this instance the order is

obligatorily VO – and it is not chance that this is the obligatory word order in Spanish. Moreover, the intonational pattern and the expression of topicality also follow the Spanish mold. Although the constituents of the VP are "German," the VP itself is not; it is in fact a "Spanish" VP.

These examples are prime instances of one of the themes that run through this book: when bilinguals code-switch, they do not simply go "back and forth" from one language to another. Nor is code-switching about inserting words of one language into the other or alternating from one language to the other. These shallow descriptions do not provide us with insight into the properties of code-switching because code-switching involves establishing a network of dependencies among the disparate constituents that conform a sentence structure. When you take a syntactic object S (of language L1) and you merge it with constituent c (of language L2), both S and c are altered. Dependencies between S and c may be built, which affect the morphosyntactic composition of both as well as the spell-out of its morphemes and prosody. Even the word order of the constituents in S may undergo a full switch as a consequence of S merging with c (and vice versa). Unfortunately, the literature has rarely focused its attention on this fact, I surmise because the separationist perspective fogs the view. Let me exemplify this with two canonical examples.

In a celebrated series of articles and books, Carol Myers-Scotton and her collaborators (Myers-Scotton 1993, 2002, Myers-Scotton and Jake 2009) have put forward a view of code-switching based on the notion of the Matrix Language Frame. Very briefly, the idea is that code-switching always involves a language that provides the function morphemes (or a relevant subset of the function morphemes, the so-called "late outsider morphemes"), creating a spine into which the speaker can insert constituents from the other language. Thus, Myers-Scotton views code-switching as uniformly insertionist, as in (2). But this view of code-switching is extremely limiting. Take, for instance, example (4). Myers-Scotton would say that Basque is the matrix language in (4a) and Spanish is in (4b). But, how do we account for the "trans-linguistic" dependency? Is the concept of matrix language enough to account for it? Or take (5). Myers-Scotton could argue that in this sentence Spanish is the matrix language and this is the reason that the German object is dislocated rather than topicalized – so far so good. However, the matrix frame model has no insight on how the Spanish clitic pronoun is in a dependency with the German object or how the gender of the clitic is chosen. Myers-Scotton's theory presents code-switching as putting pieces from two puzzles together, with very little attention being paid to the dependencies among those pieces.

No less famous is the work of Pieter Muysken (in particular, Muysken 2000). He has developed a three/four-way typology of code-switching based on the notions of *insertion* and *alternation*, terms that mean approximately what they appear to mean. Insertionist code-switching consists of putting words or

constituents of L1 into a discourse that is mostly structured as an L2 discourse, as in (3). Alternative code-switching swims between the two languages, as in (2). Again, this perspective on code-switching seems to hinder insight. To repeat: when two syntactic objects are merged together, each establishes syntactic dependencies with the other. Consider again example (5). Muysken would probably classify this example as alternational code-switching: the sentence begins in German, switches to Spanish and then back to German. Again, this gives us no insight on the appearance of the clitic, the fact that the clitic and the dislocated constituent agree in case, number, and person and, even more intriguingly, gender – the Spanish masculine is somehow able to match the neuter gender of the German constituent. Or consider example (6). Although Muysken himself provides us with (6a), he doesn't discuss why the VP has the peculiar word order that it has and how the presence of the light verb gives rise to it. I believe his separationist framework does not give him the analytical tools that would help him approach this datum.

To sum up: mainstream approaches to code-switching (and bilingual grammar in general) are fundamentally based on the idea that a bilingual has two discrete languages that are simply juxtaposed in discourse. This fundamental idea prevents scholars from training their lens on evidence that is crucial to understanding code-switching and bilingualism, including the possibility of building up dependencies in a cohesive structure.

It is not unrelated that a good chunk of bilingualism scholarship implicitly takes for granted that the notion of "grammatical system" is trivial and does not require definition or explication; it denotes the same thing for everyone and everybody recognizes what is meant when it is mentioned in an academic paper. I disagree: the notion "grammatical system" is in fact a loaded assumption, and different grammatical assumptions can lead to radically different interpretations of available data. Nonetheless, despite the lack of definition of the notion "grammatical system," it guides research questions, hypotheses, and explanations. The notion that bilinguals have "two lexicons" is also a common one and equally difficult, as will become clear throughout my discussion.

Thus, the goal of this monograph is to argue for a theory of bilingual grammar in which there is no such thing as distinct systems of linguistic competence within a person's I-language. The main data come from code-switching, but other types of data are discussed too. The focus of the analysis is the construction of linguistically cohesive language and the dependencies between the different components that constitute a sentence.

Before I start with the argument proper, I need to explain whose I-language I am actually trying to study and what data I am using. My main interest is the I-language of what I call *deep* bilinguals. Observationally speaking, deep bilinguals are people who learned two languages from birth or from a very early age (that is, they are early bilinguals) and were able to fully develop them

into adulthood. Data from deep bilinguals is the focus of the discussions on code-switching in Chapters 5 and 6. The data in these chapters come mostly from acceptability judgments obtained via surveys or in-lab experimental settings carried out in the Bilingualism Research Laboratory (BRL) at UIC. These experiments have been described in detail in published work or in dissertations that are distributed through the BRL web page, and therefore I generally omit methodological descriptions in this monograph when citing work from the lab. I also liberally use data from González-Vilbazo (2005), which involves deep bilinguals and is extracted in an experimental setting. Occasionally, I use judgments obtained informally when it was not possible for me to gather them otherwise, as long as I could contrast them with more than two consultants (see Sprouse and Almeida, 2018, which summarizes a decade of research on the replicability of acceptability judgments). I also use a good helping of data borrowed from the literature: when I do so, I describe whatever information I have regarding subjects, sources, etc.

I do not try to make any claims on later bilinguals, L2 learners, or early bilinguals whose development of the heritage language is slowed down by environmental circumstances. I am fully aware that I am leaving out data that presents its own scientific interest and eventually needs to be integrated into any theory of I-language. The reason why I focus on deep bilinguals is methodological. In his study of German/Spanish code-switching, González-Vilbazo (2005) shows that his early bilingual subjects form a coherent speech community who report similar acceptability judgments. One can conclude that those judgments reflect some property of the consultants' I-language. On the other hand, the consultants who learned German starting at age 12, although they were fully proficient, did not report consistent judgments. Thus, if we have a community of speakers S1, S2, . . . Sn and each gives different acceptability judgments on a set of stimuli, we can't really conclude anything: the differences might be due to differences in their I-languages but there could be other reasons: maybe the constructions were beyond their linguistic development in the L2; there could be the effects of "shallow processing"; disparate results can even be due to their attention limitations, the task, etc.

This difference between deep bilinguals and other bilinguals has been discussed in the literature. Toribio (2001: 215–216) provides a thoughtful discussion on the issue, but also other linguists like Zentella (1997) and Poplack (1980) report that there are substantial differences in the way that deep bilinguals and other bilinguals code-switch; only the former are able to carry out the seamless assemblage of (what appear to the outside observer as) variegated elements in one sentence or sentence fragment. These observations confirm the methodological soundness of focusing on deep bilinguals, at least while we learn more about the differences among bilinguals.

I believe that a reasonable path (not the one and only path) to learn about a person's I-language is by obtaining acceptability judgments, either in the traditional way that has been used in linguistics since time immemorial or following formal experimental protocols (see Sprouse and Almeida 2018, for a summary of arguments pro and against acceptability judgments). Regarding code-switching, one can find occasional remarks in the bilingualism literature doubting that it is appropriate to use acceptability judgments with this type of data. For instance, Muysken (2000: 13) writes that "clearly it is difficult if not impossible to rely on judgment data." Likewise, Mahootian and Santorini (1996) reject acceptability judgments in code-switching research. It seems to me that this idea must be widespread, judging from the comments that I receive whenever I submit a manuscript for publication. However, I find it disturbing: essentially, we are being told that the speech of monolingual speakers as they use their linguistic resources is rule-governed, whereas the speech of bilingual people who are also using the entirety of their linguistic resources, is not. To my mind, this is an obvious example of the monoglossic linguistic ideology that pervades the Western World – including academic specialists in the language sciences within the Western World. This monoglossic ideology takes the monolingual speech as the norm and anything bilingual as marked. It is also a beautiful example of how the separationist assumption can lead you astray: The separationist believes that a speaker has two languages in their head, each with its own set of grammatical rules. Only from this point of view can one claim that the combination of two languages gives random results. On the other hand, an integrationist scholar takes for granted that the linguistic knowledge of a bilingual person is rule-governed *in toto* and this systematic character will be reflected in their judgments of sentences, regardless of the kinds of linguistic materials that go into those sentences.

One of my language consultants, a Turkish-German bilingual, wrote the following to me (certainly tongue-in-cheek because she knew that I couldn't understand it):

(7)   Turkish/*German*
         Ama lütfen *vergessen* yapma. Biz Almancılar hepimiz *verschieden*
         konuşuyoruz. Bu *Sprache* çok flexibel. Yani hangi *Wort*'u daha önce *erinnern*
         yapıyorsan. Bu *Sprache*' de bir *Regel* yok.
         'But please don't forget. We *Almancılar* speak differently. This language is
         very flexible. Namely, the first word you remember. There are no rules in this
         language.'

While this consultant is telling me that *Almancılar* (Germans of Turkish back-ground) play around with language in an apparent chaos, she is using a tightly constructed sentence to tell me so. In this insertionist code-switched utterance, the Turkish grammatical rules are carefully respected, including the

appropriate case suffixes on the German nouns. As for the German verbs, they all appear in the appropriate German infinitival form as complements of a Turkish light verb.

Another argument that I have heard often in discussion is that since code-switched forms are stigmatized, speakers will simply give low ratings to any code-switched utterance that is presented to them (see Gullberg, Indefrey, and Muysken 2009 for discussion). I have three answers to this concern. The first is that, as Badiola, Delgado, Sande, and Stefanich (2018) show, it is indeed true that subjects with a negative attitude to code-switching give lower ratings to code-switched sentences than subjects with a positive attitude. But in the experiment that they reported, the distinctions *between* sentences were maintained proportionally regardless of attitude.[2]

A second answer to that concern is that there are methods to elicit judgments that circumvent this potential hurdle, such as two-alternative choices (Stadthagen et al. 2018). Finally, I would like to point out that if one should stop trying to access the I-language that underlies code-switching because of its marginality, then surely we should do the same to any marginalized and stigmatized language variety. I don't know any linguist who would like to walk down that path.

There are of course other linguists who cast doubt on acceptability judgments in monolingual work too, as a matter of principle (Silva-Corvalán 2001). I find this rejection surprising, since most of what we know about any human language – I would say, 99 percent – was learned by introspection and prodding the judgments of native speakers. I have yet to see a descriptive grammar of a language that is not entirely built on native speaker intuitions. The knowledge that we have been able to obtain in recent years via corpora studies or psycholinguistic work is very valuable and surely should be integrated into grammatical theory – but it is built on a foundation of native speaker judgments because only these judgments tell the researcher what to look for.

The data in Chapters 7 and 8 is exclusively based on the published literature and, as far as I can tell, it has been obtained through generally accepted protocols in psycholinguistic research. More often than not, it was impossible for me to see what kinds of bilinguals are discussed. As for Chapter 9, the data is also from published sources and obtained in a variety of ways.

The rest of this monograph is organized as follows. In Chapter 2, I continue the discussion of the separationist hypothesis and I present what is, in my view, the most formally rigorous model of separationist bilingual grammar: MacSwan's (1999, 2000) minimalist approach. This chapter also includes a brief discussion of Tom Roeper's Multiple Grammar Theory. In Chapter 3, I introduce some concepts of minimalism and distributed morphology that give theoretical shape to my proposals. This chapter also includes some